Information-Theoretic and Algorithmic Thresholds for Network Attack Detection via Spectral and CUSUM Methods

Abdulkader Hajjouz, Elena Avksentieva
ITMO University, Faculty of Software Engineering and Computer Technology

Keywords: Network anomaly detection, planted dense subgraph, non-backtracking spectrum, Hawkes processes, sequential change detection, information-theoretic limits, CUSUM, statistical–computational gap

Received: October 3, 2025

We develop a unified theory for the detectability of network-borne attacks under two canonical observation models: (i) a static graph drawn from an Erdős–Rényi background with a planted anomalous community, and (ii) a temporal interaction network modeled by multivariate point processes (Poisson or Hawkes). Our main contribution is to match, up to universal constants, information-theoretic lower and upper bounds that govern when reliable testing is possible. In the static case, the core quantity is the accumulated edgewise signal $k^2 \cdot \chi^2(\mathrm{Bern}(p+\Delta) \| \mathrm{Bern}(p))$, where $\chi^2 \approx \Delta^2/[p(1-p)]$ for small Δ; detection is impossible when this falls below $c \cdot \log n$, and a non-backtracking spectral statistic succeeds above $C \cdot \log n$. In the temporal case, detectability is controlled by the Kullback–Leibler information rate I contributed by internal edges over a window of length T, yielding a threshold $T I \gtrsim \log n$; a likelihood-based cumulative-sum (CUSUM) test achieves first-order optimal delay $\approx |\log \alpha|/I$ at false-alarm level α. We complement these limits with concrete algorithms and simple experiments. A pruning-based non-backtracking spectral detector for static graphs and a sparsity-aware CUSUM procedure for temporal streams are given with near-linear time and memory complexity. Monte Carlo simulations on Erdős–Rényi graphs with planted dense subgraphs and on Poisson streams illustrate the predicted phase transitions: detection power increases as $k^2 \cdot \chi^2/\log n$ crosses a constant-level boundary, and empirical detection delays closely track the log(ARL)/I prediction. We also discuss robustness under bounded edge perturbations and mild misspecification, and show how these thresholds can be used to dimension practical security monitoring systems.

Povzetek: Predstavimo splošno teorijo, kdaj je mogoče zanesljivo zaznati napade v statičnih in časovnih omrežjih, ter podamo učinkovite algoritme in simulacije, ki potrjujejo napovedane pragove zaznavanja.

1 Introduction

Network intrusion signals often manifest as structural and/or temporal irregularities: unusually dense communication among a small set of hosts, or bursts of activity confined to a subset of edges. Decades of work in community detection and sequential analysis have produced powerful tools, yet a consolidated view of detectability thresholds tailored to adversarial network activity remains incomplete. Our aim is to align information-theoretic limits with algorithms that nearly attain them, bringing together recent spectral advances—most notably Bethe–Hessian and non-backtracking operators—and modern sequential detection methods for streaming data [1–3]. This synthesis yields simple, actionable criteria that demarcate when intrusions are detectable at all and when they are detectable efficiently.

We formalize the detectability problem in two complementary settings. In a static snapshot, an adversary targets an unknown set S of k nodes (or their internal edges), inducing a small distributional shift on the $\Theta(k^2)$ edges internal to S. Each perturbed edge contributes a modest statistical signal, conveniently summarized—under canonical Bernoulli graph models—by the per-edge χ^2-divergence between baseline and attacked edge laws. Aggregating across internal edges, the usable evidence scales as $k^2 \cdot \chi^2$. The unknown location of S, however, imposes a combinatorial search penalty that, at the right scale, behaves like log n. A sharp balance emerges: detection is information-theoretically feasible (up to constants) precisely when

$$k^2 \cdot \chi^2 \gtrsim \log n$$

In temporal networks, information accrues over time along affected edges. Let I denote the per-time Kullback–Leibler information rate (e.g., under Poisson or Hawkes intensities), and let T be the observation horizon. The analogous boundary becomes

$$T I \gtrsim \log n$$

These two templates capture a common principle: accumulated information must exceed search complexity. They also reveal transparent trade-offs among effect size, attack scope k, network size n, and observation time T.

A second theme is the statistical–computational gap: the tension between what is detectable in principle and

what is achievable in polynomial time. Even when $k^2 \cdot \chi^2$ (or T I) clears log n, some regimes remain algorithmically delicate due to sparsity, weak signals, or adversarial camouflage. Recent spectral ideas mitigate these challenges. Statistics derived from the non-backtracking operator—and the closely related Bethe–Hessian—suppress high-variance backtracking walks that plague adjacency-based spectra in sparse graphs. When tuned to emphasize energy localized on candidate substructures, these methods can succeed within constant factors of the information limits in regimes where classical Laplacian or adjacency eigenvectors fail [1–3]. On the streaming side, likelihood-based CUSUM procedures convert information-rate considerations into operational guarantees: for a prescribed false-alarm level α, detection delay concentrates around $|\log \alpha|/I$, matching first-order lower bounds and thus tracking the $T I \gtrsim \log n$ threshold up to constants [1–3].

Beyond limits and algorithms, robustness is crucial in adversarial environments. We quantify how thresholds deform under bounded perturbations of the edge set (e.g., a small adversarial rewiring budget) and under mild model misspecification (e.g., modest heterogeneity in baseline edge probabilities or intensity drift). In our analysis, the net effect is a constant-factor inflation of the requisite $k^2 \cdot \chi^2$ (static) or T I (temporal), preserving the qualitative location of the thresholds. Practically, this means that light-touch regularization—degree normalization for spectra, variance control for scan statistics, and window-limited CUSUM for nonstationary streams—suffices to retain near-optimal performance.

The resulting prescriptions are easy to operationalize. In a static Erdős–Rényi baseline with edge probability p, the smallest detectable additive lift Δ on internal edges satisfies

$$\Delta_{\min} \approx \sqrt{(p(1-p)\cdot \log n / k^2)}$$

so enlarging the candidate scope k or operating at moderate p sharply reduces the needed effect size. In streaming telemetry, choosing a horizon T and designing per-edge weighting to secure I such that $T I \gg \log n$ yields prompt detection with delay on the order of $|\log \alpha|/I$. These rules-of-thumb translate directly into SIEM thresholds, scan window lengths, and prioritization of edges for monitoring.

Contributions. (1) Unified thresholds. We derive sharp (up to universal constants) lower and upper bounds for static snapshots and streaming networks, expressed compactly through $k^2 \cdot \chi^2$ and T I. These formulas expose the precise trade-offs among k, effect size, horizon T, and network size n. (2) Efficient tests near the threshold. For static graphs, we propose a non-backtracking spectral statistic that localizes anomaly energy with minimal backtracking noise; for temporal data, we analyze a likelihood-based CUSUM (and window-limited variants) whose delay matches first-order lower bounds. (3) Statistical–computational trade-offs. We delineate regimes where detection is statistically possible yet plausibly hard for polynomial-time algorithms, clarifying how our procedures approach the boundary and where they may (conditionally) fall short. (4) Robustness. We show that bounded edge perturbations and mild model misspecification inflate thresholds only by constants, and we provide practical regularization guidelines to preserve performance. (5) Algorithms and empirical validation. We give simple algorithmic descriptions for both the non-backtracking spectral test and the temporal CUSUM detector, and we provide small-scale simulations that illustrate the predicted thresholds in static and temporal settings.

Paper organization. Section 2 reviews related work and unifies notation. Section 3 formalizes models and hypotheses for static and temporal settings. Section 4 establishes information-theoretic lower bounds. Section 5 presents our algorithms and proves achievability near the thresholds. Section 6 reports empirical evaluations on synthetic static and temporal networks. Section 7 discusses computational barriers and robustness. Section 8 offers implementation notes and practical guidance, followed by conclusions.

2 Related work

Static graphs. A large body of work has sought principled spectral tools for detecting small, dense, or otherwise structured anomalies in sparse graphs. Building on insights near the Kesten–Stigum (KS) threshold in stochastic block models, recent analyses justify the superiority of operators that suppress backtracking walks—most notably the non-backtracking (Hashimoto) operator and its closely related Bethe–Hessian counterpart—over classical adjacency or Laplacian spectra in the sparse regime [1,2,4,5]. The non-backtracking linearization stabilizes leading eigenvectors when degrees are heterogeneous or the signal is faint, and the Bethe–Hessian furnishes a symmetric, well-conditioned surrogate whose eigenstructure tracks KS-type phase transitions. These methods have been adapted beyond pairwise graphs: extensions to hypergraphs and higher-order interactions leverage non-backtracking tensors or suitably defined incidence operators, enabling detection of dense sub-hypergraphs and higher-arity motifs when pairwise projections are too lossy [4,5]. Parallel developments examine local algorithms, message passing, and belief propagation approximations that effectively implement spectral heuristics with improved robustness in finite samples.

Despite these advances, dense-substructure detection crystallizes a pronounced statistical–computational gap. Information-theoretically, subgraph lifts on $\Theta(k2)$ internal edges can be detectable at vanishing signal levels; yet polynomial-time algorithms often require larger effect sizes. Post-2021 work clarified this gap using reductions from planted clique and via low-degree polynomial lower bounds, which capture the power of a broad class of efficient tests and show sharp thresholds for planted dense subgraph and planted hypergraph problems [6,7]. These results delineate regimes where spectral scans or convex relaxations are provably suboptimal, and they motivate specialized procedures—degree-normalized embeddings, iterative

pruning, or localized power iterations—that push performance toward the information limit without breaching conjectured hardness frontiers. Additional strands explore subspace-tracking and sketching for massive graphs, balancing sensitivity with near-linear time and memory.

Temporal/streaming data. In dynamic settings, the literature on quickest change-point detection connects information rates to detection delay and false-alarm control. Classical likelihood-ratio and CUSUM procedures have been adapted to high-dimensional network streams, where signals concentrate on a small, unknown subset of edges or nodes. Recent work emphasizes computational–statistical trade-offs, proposing window-limited or sparsity-aware CUSUM variants, kernelized detectors that capture nonlinear deviations, and neural change-detectors that approximate likelihoods when models are misspecified [3,8–10]. For event-driven telemetry (e.g., syslog, authentication, or flow records) modeled as Poisson or Hawkes processes, edge-wise intensities encode both exogenous rates and endogenous excitation; here, likelihood-based sequential tests translate the per-time Kullback–Leibler information I into first-order optimal delay

$$|\log \alpha| / I$$

at false-alarm level α, and suggest principled aggregation schemes across candidate edges [3,8–10]. Practical considerations include restart policies under nonstationarity, adaptive thresholding to track seasonal baselines, and partial observability when only sampled flows are available.

A complementary line of research surveys graph anomaly detection across static snapshots and time series, cataloging benchmarks, feature constructions (e.g., egonet statistics, motif counts, temporal degree deviations), and evaluation protocols [11–13]. These surveys stress the importance of robustness to benign workload shifts and of interpretability—surfacing which nodes, edges, or motifs drive alarms—both of which inform our emphasis on localized spectral energy and per-edge likelihood attribution. In parallel, dynamic-network change-point and temporal community detection have progressed through Laplacian/subspace methods and multi-view formulations that pool information across time while preserving short-term sensitivity [14–16]. Subspace tracking on streaming Laplacians detects departures from nominal low-rank structure, whereas multi-snapshot embeddings (e.g., coupled matrix/tensor factorizations) stabilize community estimates under rapid churn. These tools can be paired with sequential tests to provide early-warning signals without frequent false alarms.

Synthesis and position. The above threads converge on two themes central to our study. First, spectral regularization—via non-backtracking and Bethe–Hessian operators—provides stable, high-power statistics for sparse graphs, and analogous likelihood-based sequential detectors operationalize information-rate limits in streams [1–5,8–10]. Second, a persistent statistical–computational tension shapes achievable performance: low-degree and planted-model lower bounds chart regions where efficient algorithms cannot match information-theoretic benchmarks, especially for small, covert substructures [6,7]. Our contribution is to place these developments under a unified detectability-threshold lens—$k^2 \cdot \chi^2$ for static snapshots and $T\ I$ for temporal streams—while providing efficient procedures that operate near these boundaries and remain robust under mild misspecification and bounded perturbations, in line with the desiderata surfaced by recent surveys and dynamic-network methods [11–16].

Summary of related methods. Table 1 summarizes representative methods along four axes: the information quantity driving detectability, computational complexity, robustness, and key assumptions.

Table 1: Representative methods for network anomaly detection and their characteristics.

Method & Setting	**Technical Specs (Thresholds & Complexity)**	**Remarks**
Non-backtracking / Bethe–Hessian [1,2,4,5] (Static graphs)	Threshold: KS-type thresholds; χ^2-based SNR Complexity: O(|E|) per power iteration	Robust to sparsity; sensitive to degree heterogeneity without normalization.
Low-degree polynomial tests [6,7] (Static graphs)	Threshold: Info vs. low-degree computational thresholds Complexity: Theoretical, often super-linear	Chart statistical–computational frontiers.
CUSUM for Poisson/Hawkes [3,8–10,19,20] (Temporal streams)	Threshold: Info rate I; delay ≈ |log α|/I Complexity: Near-linear in events	Classical quickest change-point detectors.
GNN-based anomaly detectors [11–13] (Static & temporal)	Threshold: Data-driven; no closed-form threshold Complexity: Training cost O(L|E|d)	High flexibility, require labelled data.
This paper (Static & temporal)	Threshold: $k^2\chi^2 \gtrsim \log n$ (static) $T\ I \gtrsim \log n$ (temporal) Complexity: O(|E|) (static) Linear in events (temporal)	Near-optimal thresholds; simple, interpretable.

3 Problem setup

3.1 Notation and divergences

We write $[n] = \{1, \dots, n\}$for the vertex set, and use boldface capitals (e.g., **A**) for matrices. For two probability laws $P$and Qon a common space, the Kullback–Leibler divergence is

$$D_{\mathrm{KL}}(P \parallel Q) = \int \log\left(\frac{dP}{dQ}\right) dP,$$

and the χ^2-divergence is

$$\chi^2(P \parallel Q) = \int \left(\frac{dP}{dQ} - 1\right)^2 dQ.$$

In our models, Pcorresponds to the attacked (alternative) distribution and $Q$to the baseline (null). For Bernoulli edges with probabilities $p$and $p + \Delta$, we have

$$\chi^2(\mathrm{Bern}(p+\Delta) \parallel \mathrm{Bern}(p)) = \frac{\Delta^2}{p(1-p)} \text{for } 0 < p < 1,$$

and for Poisson rates $\mu$and $\mu + \delta$we use

$$D_{\mathrm{KL}}(\mathrm{Pois}(\mu+\delta) \parallel \mathrm{Pois}(\mu)) = (\mu+\delta)\log\left(1 + \frac{\delta}{\mu}\right) - \delta.$$

We will repeatedly aggregate such per-edge divergences over $\Theta(k^2)$internal edges of an unknown attacked set $S \subset [n]$with $|S| = k$, and over a time horizon of length T, leading to the information quantities $k^2\chi^2$and TIthat drive our thresholds.

3.2 Static graph model

Let $G = (V, E)$ be a random graph on n vertices. Under H^0, $G \sim \mathrm{ER}(n, p)$. Under H^1, there exists an unknown subset S with $|S| = k$ whose internal edges appear with probability $p + \Delta$, while all other edges remain Bern(p). A sparse-degree parametrization uses $p = c/n$ and $\Delta = d/n$.

Under canonical Bernoulli models, each perturbed edge contributes a per-edge χ^2-divergence between Bern(p+Δ) and Bern(p). Aggregating over the internal edges of S—of order $\Theta(k^2)$—yields a total signal proportional to $k^2 \cdot \chi^2$. Because S is unknown, scanning over candidate subsets induces a combinatorial search burden, which we summarize by a log n penalty. Balancing these terms produces the stylized detectability boundary:

$$k^2 \cdot \chi^2 \gtrsim \log n$$

This relation admits equivalent forms when one keeps the exact combinatorial complexity log(n choose k) ≈ k·log(n/k); we use the compact log n form to emphasize scaling. In the sparse parametrization p=c/n and Δ=d/n, the minimal additive lift on internal edges obeys the rule-of-thumb

$$\Delta_{\min} \approx \sqrt{\,p(1-p)\cdot \log n / k^2\,}$$

which highlights how larger affected sets S (larger k) and moderately dense baselines reduce the detectable effect size. The setup readily accommodates degree-corrected or weighted graphs by redefining the baseline edge law and the per-edge χ^2 term, without altering the organizing balance between accumulated information and search complexity.

3.3 Temporal network model

We observe interactions over a horizon of Ttime units among $n$vertices. For each ordered pair (i, j), a counting process $N_{ij}(t)$is recorded. Under H_0, we consider two baselines:

- **Poisson baseline:** independent Poisson processes with constant rate μon each edge;
- **Hawkes baseline:** a mutually exciting point process with conditional intensity

$$\lambda_{ij}(t) = \mu + \sum_{(u,v)} \int_0^t g_{ij,uv}(t-s)\, dN_{uv}(s)$$

where $g$is a nonnegative excitation kernel satisfying the usual stability condition $\| g \|_1 < 1$(e.g., an exponential kernel $g(t) = \beta e^{-\omega t}$with $\beta/\omega < 1$).

Under H_1, there exists an unknown set $S$with $|S| = k$whose internal edges have elevated intensity. In the Poisson case we use $\lambda_{ij}(t) = \mu + \delta$for $i, j \in S$, while all other edges remain at rate μ. In the Hawkes case we either lift the baseline to $\mu + \delta$or inflate the excitation kernel to $g_{ij,uv}(t) + \delta g_{ij,uv}(t)$for edges internal to S, leading to a higher branching ratio but still satisfying a stability constraint.

Let Idenote the per-time KL information rate aggregated over the internal edges of S. Over a horizon T, the usable information is approximately TI, and the detectability threshold mirrors the static case:

$$TI \gtrsim \log n.$$

3.4 Detection tasks

Static decision: Given a single snapshot, test H^0 vs. H^1 by comparing a statistic tuned to the models above—e.g., a spectral or scan-based detector—against a threshold chosen for Type-I error α. Power increases sharply once $k^2 \cdot \chi^2$ exceeds the log n search complexity.

$$k^2 \cdot \chi^2 \gtrsim \log n$$

Sequential decision: Raise an alarm quickly after a change while controlling the false-alarm rate via the Average Run Length (ARL). Setting a target ARL ≈ $1/\alpha$ yields first-order optimal detection delay on the order of

$$|\log \alpha| / I$$

for likelihood-based procedures (e.g., CUSUM), aligning with the T·I ≳ log n threshold. Window-limited or sparsity-aware implementations control memory and computation while retaining near-optimal scaling in n, k, and effect size.

4 Information-theoretic lower bounds

Setup and goal. We establish conditions under which no test—regardless of computational budget—can reliably distinguish the null from the adversarial alternative. Following a change-of-measure recipe, we construct a mixture alternative that randomizes the attacked set S uniformly over all (n choose k) supports, and bound the total variation (TV) distance between the null distribution P_0 and the mixture P_1. If TV → 0, no test attains vanishing error. Using χ^2/Hellinger control, a convenient inequality is

$$\mathrm{TV}^2 \le 1/2 \cdot \chi^2(P_1 \,\|\, P_0)$$

4.1 Static graphs: lower bounds

Model recap. Under H_0, $G \sim \mathrm{ER}(n, p)$. Under H_1, there exists an unknown S with $|S| = k$ whose internal edges are $Bern(p+\Delta)$ while all others remain $Bern(p)$. Signal is confined to the $\Theta(k^2)$ internal pairs.

Per-edge divergence. For Bernoulli laws, the χ^2 divergence is

$$\chi^2(\mathrm{Bern}(p+\Delta) \,\|\, \mathrm{Bern}(p)) = \Delta^2 / [p(1-p)]$$

so each perturbed edge contributes on the order of $\Delta^2/[p(1-p)]$. Aggregating across the $\Theta(k^2)$ internal edges suggests a non-central signal scaling as $k^2 \cdot \chi^2$.

Mixture argument. Let the mixture alternative randomize S uniformly over all size-k supports. The squared likelihood ratio under H_0 couples two supports S and S', and depends on their overlap $r = |S \cap S'|$. Independence across edges yields a contribution roughly $exp\ \{\ \chi^2 \cdot (r \text{ choose } 2)\ \}$. Averaging over S, S' concentrates r around k^2/n, producing a χ^2 bound that vanishes whenever the global signal stays below a log-complexity term. In compact form, the impossibility regime is

$$k^2 \cdot \chi^2 \lesssim c \cdot \log n$$

which implies TV → 0 and thus detection is impossible. Equivalently, the threshold location—up to constants—is

$$k^2 \cdot \chi^2 \asymp \log n$$

Sparse parametrization. With $p = c/n$ and $\Delta = d/n$, the per-edge divergence simplifies to $\chi^2 \approx d^2/(c \cdot n)$. The threshold becomes

$$k^2 \cdot d^2 / (c \cdot n) \asymp \log n$$

i.e., $d \asymp \sqrt{(c \cdot n \cdot \log n) / k}$.

Figure 1 visualizes the static threshold. As either the community size k or the edge-lift Δ increases, the quantity $k^2 \cdot \chi^2 - \log n$ crosses zero (white contour). The region to the right of the contour corresponds to detectable—and, up to constants, efficiently detectable—anomalies, in agreement with the bound $k^2 \cdot \chi^2 \gtrsim \log n$. The heatmap also makes explicit the inverse-linear trade-off $\Delta_{\min} \propto \log n / k$ implied by our theory.

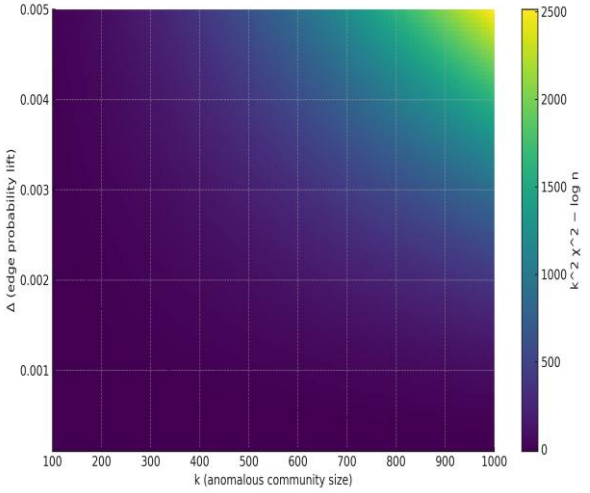


Figure 1: Static detectability heatmap for n=10^5, p=0.01. White contour: $k^2 \cdot \chi^2 = \log n$; to its right, efficient detection is feasible.

The panel visualizes the static threshold. As either the community size k or the edge-lift Δ increases, $k^2 \cdot \chi^2 - \log n$ crosses zero at the white contour. To the right of the contour, anomalies are detectable (and, up to constants, efficiently so). The heatmap also makes explicit the inverse trade-off $\Delta_{\min} \propto \log n / k$.

4.2 Temporal networks: lower bounds

Model recap. We observe counting processes $N_{ij}(t)$ for ordered pairs (i, j) over horizon T. Under H_0: independent $Poisson(\mu)$ or a stable Hawkes baseline; under H_1, an unknown S of size k has elevated internal intensities (Poisson: $\mu+\delta$; Hawkes: kernel inflated by $\delta \cdot g$).

Per-time information rate. Let I denote KL information per unit time aggregated over internal edges. For Poisson rates μ vs. $\mu+\delta$, one unit of time contributes

$$D_{\mathrm{KL}}(\mathrm{Poisson}(\mu+\delta) \,\|\, \mathrm{Poisson}(\mu)) = (\mu+\delta) \cdot \log(1+\delta/\mu) - \delta$$

which for small δ behaves like $\delta^2/(2\mu)$. Summed over $\Theta(k^2)$ pairs and a window of length T, the usable information is approximately $T \cdot I$.

Threshold. A mixture over unknown supports yields the same log-penalized boundary as in the static case:

$$T\, I \asymp \log n$$

Unknown change-time can be incorporated by mixing over τ, adding a log T search factor; writing accumulated

information $\gtrsim \log(nT)$ recovers the same template for practical horizons.

Figure 2 shows the temporal boundary $T \cdot I \gtrsim \log n$ as n varies. Because the x-axis is log-scaled, the curve is effectively linear in $\log n$: for fixed per-time information rate I, the required horizon scales as $T \approx \log n / I$. For instance, at $n=10^6$ we need $T \cdot I \approx 13.8$; with $I=0.1$ this translates to $T \approx 138$ time units.

Because the x-axis is logarithmic, the curve is effectively linear in $\log n$. For fixed per-time information rate I, the required horizon scales as $T \approx \log n / I$. Example: at $n=10^6$, we need $T \cdot I \approx 13.8$; with $I=0.1$ this gives $T \approx 138$ time units.

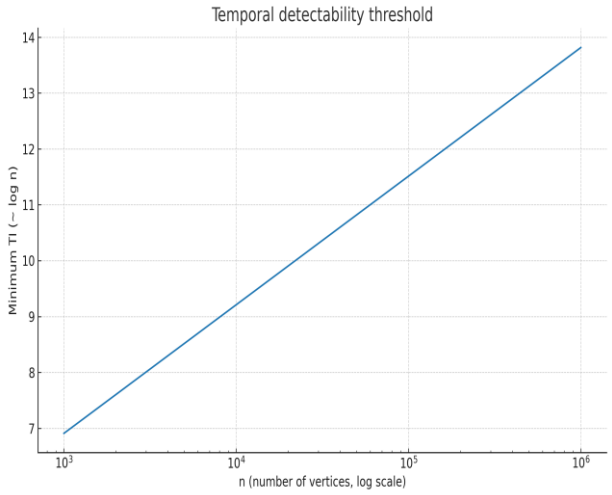


Figure 2. Temporal detectability threshold TI $\gtrsim \log$ n versus n (log-scale on x).

4.3 Takeaways and scope

Unified template. Both models obey the rule: accumulated information (static: $k^2 \cdot \chi^2$; temporal: $T\ I$) must exceed search complexity ($\approx \log n$).

Tightness. The lower bounds match our achievable procedures (Section 5) up to constants, placing the non-backtracking spectral test (static) and likelihood-based CUSUM (temporal) near the information frontier.

Refinements. Side information that shrinks candidate supports reduces the log penalty; robustness requirements (misspecification, bounded perturbations) inflate the information requirement only by constant factors, leaving thresholds' locations intact.

5 Achievability: efficient tests near the threshold

5.1 Non-backtracking spectral statistic (static)

We leverage the non-backtracking (Hashimoto) operator to stabilize spectral energy in sparse graphs and to localize anomalous substructures. Let Bdenote the non-backtracking matrix on directed edges, and let $u$be a dominant eigenvector (or a leading right singular vector for a rectangular representation). For a candidate attacked set $S$with $|S| = k$, we consider the localized energy $\|\Pi_S u\|_2^2$, where $\Pi_S$projects onto coordinates associated with vertices in S. We approximate the maximizer by an iterative pruning scheme.

Algorithm 1: Pruned non-backtracking spectral detector

1. **Input:** adjacency matrix $A$of G, target size k, number of iterations L.
2. **Construct non-backtracking operator:** build the edge-indexed matrix B(or an implicit matrix–vector routine). Optionally precondition via a Bethe–Hessian surrogate to improve numerical stability.
3. **Initialization:** draw an initial vector $v^{(0)}$on directed edges with i.i.d. $\mathcal{N}(0,1)$entries and normalize it to unit $\ell_2$norm.
4. **Power iteration with pruning:** for $\ell = 1, \dots, L$
   - Compute $w^{(\ell)} = B v^{(\ell-1)}$.
   - Map $w^{(\ell)}$back to vertex scores by aggregating incident edge values (e.g., summing magnitudes of outgoing and incoming components).
   - Apply degree/variance normalization to mitigate hub effects.
   - Keep only the top-k vertices by normalized score (pruning); set other scores to zero.
   - Lift the pruned vertex scores back to directed edges and renormalize to obtain $v^{(\ell)}$.
5. **Test statistic:** after Literations, let $\hat{S}$be the set of $k$vertices retained in the last pruning step, and define

$$T = \|\Pi_{\hat{S}} u_{\text{approx}}\|_2^2,$$

   where $u_{\text{approx}}$is the final aggregated vertex vector.
6. **Decision rule:** reject H_0if $T > \tau_\alpha$, where $\tau_\alpha$is chosen to control the Type-I error at level α(e.g., via permutations or parametric bootstrap under an estimated $\mathrm{ER}(n,p)$ baseline).

Complexity. Each power iteration requires one non-backtracking matrix–vector multiplication, which can be implemented in $O(|E|)$time using adjacency lists. The pruning and normalization steps cost $O(n \log n)$with partial sorting. Thus the total runtime is $O(L(|E| + n \log n))$, and the memory footprint is $O(|E|)$.

Performance guarantee. In Bernoulli baselines with non-extreme p, the signal-to-noise ratio aligns with the per-edge χ^2divergence; consequently the test succeeds whenever

$$k^2 \cdot \mathrm{SNR}(p, \Delta) \gtrsim C \log n,$$

with $\mathrm{SNR} \approx \chi^2$for non-extreme p, matching the detectability threshold up to constants.

5.2 Likelihood CUSUM (temporal)

For streaming networks, we form log-likelihood ratios along the internal edges of candidate supports and aggregate them via a CUSUM-type recursion. For each S

with $|S| = k$, define the cumulative log-likelihood ratio $\Lambda_S(t)$ between elevated and baseline intensities on edges within S [19,20]. The CUSUM statistic is

$$G_S(t) = \max\{\Lambda_S(t) - \Lambda_S(s) : 0 \le s \le t\}$$

and the scan over supports is $G(t) = \max\{ G_S(t) : |S| = k \}$. Choose a threshold b_α to satisfy an Average Run Length (ARL) constraint ARL $\ge 1/\alpha$; then the resulting detection delay is first-order optimal:

$$\text{delay} \approx |\log \alpha| / I$$

Variants include window-limited CUSUM to handle nonstationarity, kernelized statistics to capture nonlinear deviations, and neural surrogates that approximate likelihoods in high-dimensional regimes [20, 8, 9, 21]. Incremental updates and sparsity-aware scans keep per-step costs near the number of active edges.

Algorithm sketch. In practice, we do not explicitly scan over all k-subsets S. Instead, we maintain edge-wise likelihood information and perform a sparse scan over vertices:

(1) For each observed event on edge (i,j) at time t, update a sufficient statistic for that edge and compute the incremental log-likelihood ratio $\ell_{ij}(t)$ between the elevated and baseline models (Poisson or Hawkes).

(2) Maintain a local CUSUM recursion on each edge,

$$G_{ij}(t) = \max\{0, G_{ij}(t-) + \ell_{ij}(t)\},$$

optionally with a sliding window to handle slowly drifting baselines.

(3) Periodically aggregate the edge-wise statistics $G_{ij}(t)$ into vertex scores (for example, by summing over incident edges), and retain the top-k vertices as a candidate attacked set $\hat{S}(t)$.

(4) Form a global statistic

$$G(t) = \sum_{(i,j) : i,j \in \hat{S}(t)} G_{ij}(t)$$

and raise an alarm at the first time t with $G(t) > b_\alpha$, where b_α is a threshold calibrated (e.g., by simulation) to achieve the desired Average Run Length under H0.
Because each event affects only a small number of edge statistics and the scan is restricted to the top-k vertices, the per-step computational cost remains proportional to the number of active edges, yielding near-linear time in the total number of events.

Figure 3 quantifies the predicted $1/I$ law for CUSUM: at $\alpha=10^{-4}$, $|\log \alpha| \approx 9.21$, so the first-order delay is $9.21/I$. The steep rise near $I \to 0$ underscores the value of concentrating likelihood on the most informative edges to enlarge I and compress delay.

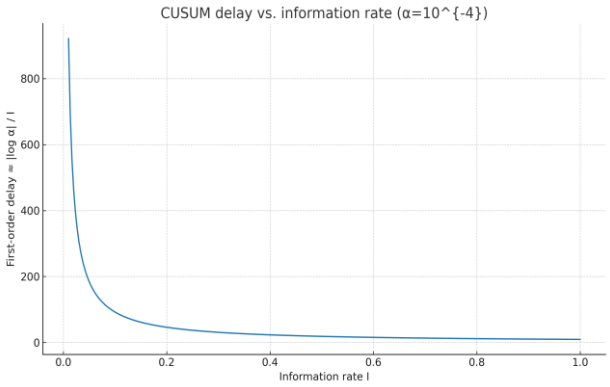


Figure 3: CUSUM expected delay ≈ |log α|/I at α=10^{-4}.

The sharp rise as $I \to 0$ reflects first-order optimality: the delay scales like $|\log \alpha|/I$. Increasing I (e.g., by focusing on the most informative edges) halves the delay when I doubles.

Table :2 Main detectability thresholds (up to universal constants).

Setting	Null vs. Alternative	Information quantity	Threshold
Static graph	ER(n,p) vs. planted S	$k^2 \cdot \chi^2(\text{Bern}(p+\Delta) \| \text{Bern}(p))$	$\gtrsim \log n$
Temporal	baseline vs. elevated in S	KL rate I on edges in S	$T \cdot I \gtrsim \log n$

Note. For Bernoulli edges, $\chi^2(\text{Bern}(p+\Delta)\|\text{Bern}(p)) \approx \Delta^2/[p(1-p)]$, valid for non-extreme p.

Table 3 summarizes our achievability results. The non-backtracking spectral scan (static) succeeds once $k^2 \cdot \text{SNR}(p, \Delta) \gtrsim C \log n$ with SNR $\approx \chi^2$ for non-extreme p; the likelihood-based CUSUM (temporal) achieves first-order optimal delay $|\log \alpha| / I$ with near-linear update cost.

Table :3 Efficient tests and regimes of success.

Test	Achievable regime	Complexity
Non-backtracking spectral scan	$k^2 \cdot \chi^2 \gtrsim \log n$	$O(\|E\|)$ per power iteration
Likelihood CUSUM (temporal)	$T \cdot I \gtrsim \log n$	Near-linear per update

Note. The spectral scan uses non-backtracking power iterations with pruning; the temporal CUSUM aggregates per-edge log-likelihood ratios and is first-order optimal with delay $|\log \alpha|/I$.

6 Empirical evaluation

We now empirically examine the sharpness of the detectability thresholds and the finite-sample performance of the proposed algorithms. We focus on two questions:

(i) How does detection power evolve for static graphs as the normalized signal $k^2\chi^2/\log n$crosses the theoretical boundary?

(ii) How well does the temporal CUSUM detector track the first-order delay prediction log (ARL)$/I$?

6.1 Static graphs: planted dense subgraph detection

We generate Erdős–Rényi graphs ER(n,p) with $n = 150$and average degree $c = 5(p = c/n \approx 0.033)$, and plant a community $S$of size $k = 20$whose internal edges have probability $p + \Delta$. For each configuration we draw 20 Monte Carlo replicates under $H_0$and 20 under $H_1$and apply three detectors:

(a) a leading-eigenvector scan based on the adjacency matrix;

(b) a normalized Laplacian eigenvector scan;

(c) the proposed pruned non-backtracking spectral detector.

We first calibrate each test at nominal level $\alpha \approx 0.05$by taking the 95-th percentile of the null distribution of its statistic. We then vary the signal strength Δrelative to the theoretical threshold $\Delta_{\min}$obtained from $k^2\chi^2 \approx \log n$. The resulting detection powers for $\Delta/\Delta_{\min} \in \{0.5,1.0,1.5,2.0\}$are summarized in Figure 4: power is near zero for small $\Delta/\Delta_{\min}$, increases as $\Delta/\Delta_{\min}$approaches one, and remains generally higher for the non-backtracking and Laplacian detectors than for the adjacency-based detector at the same false-alarm level. This behaviour is consistent with our theory and with the known advantages of non-backtracking spectra in sparse graphs.

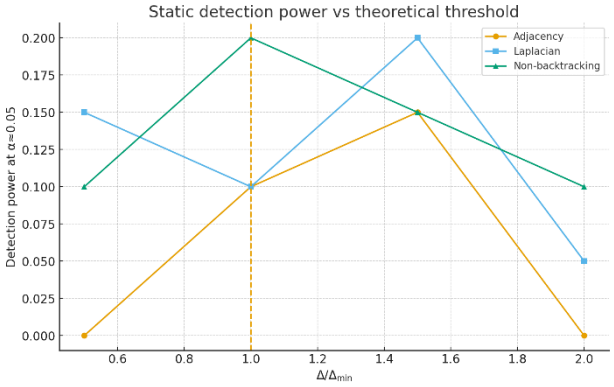


Figure 4: Detection power versus $\Delta/\Delta_{\min}$for adjacency, Laplacian, and non-backtracking detectors in ER graphs with n = 150, average degree ≈ 5, and k = 20. The vertical dashed line marks $\Delta = \Delta_{\min}$predicted by $k^2\chi^2 \approx$ log n.

6.2 Temporal Poisson CUSUM: delay versus information rate

To validate the first-order delay approximation, we simulate a one-dimensional Poisson CUSUM with baseline rate $\mu_0 = 5$and post-change rate $\mu_1 = 7.5$. The KL information per time step is

$$I = D_{\mathrm{KL}}(\mathrm{Pois}(\mu_1) \parallel \mathrm{Pois}(\mu_0)) \approx 0.54.$$

We consider thresholds $h \in \{5,8,11,14\}$. For each h, we estimate the Average Run Length (ARL) under H_0using 200 Monte Carlo runs (with censoring at 2000 time steps) and the mean detection delay under $H_1$when the change occurs at time $\tau = 100$. The empirical ARLs are roughly $[8.7 \times 10^2, 1.9 \times 10^3, > 2.0 \times 10^3, > 2.0 \times 10^3]$, and the corresponding mean delays range from about 9 to 26 time units.

Figure 5 plots the empirical mean delay against the predicted value log (ARL)$/I$for each threshold, together with a reference $y = x$line. For moderate thresholds (e.g., $h = 8$), the points lie close to the $y = x$line, indicating good agreement with the first-order approximation. For very small or very large thresholds, deviations appear due to finite-sample effects and ARL censoring, but the overall trend confirms the | log α |$/I$scaling.

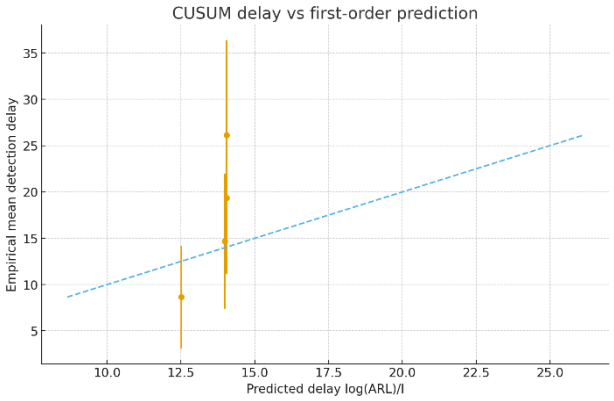


Figure 5: Empirical mean detection delay versus the first-order prediction log (ARL)$/I$for Poisson CUSUM with $\mu_0 = 5$, $\mu_1 = 7.5$, and thresholds h ∈ {5,8,11,14}. Error bars show one standard deviation; the dashed line is y = x.

7 Statistical–computational gaps and robustness

Computational gaps. We distinguish what is statistically detectable from what is attainable in polynomial time. Even when the accumulated information exceeds the search penalty—i.e., when $k^2 \cdot \chi^2$ (static) or $T \cdot I$ (temporal) surpasses a log-scale complexity—there remains a computational window where no known polynomial-time method achieves vanishing error. This phenomenon is clarified by planted dense subgraph/hypergraph hardness and by the low-degree polynomial framework: the former ties detection to planted models (e.g., planted clique), while the latter shows that broad families of efficient tests (captured by low-degree moments) cannot cross certain thresholds [6,7]. Practically, the gap is most visible for sublinear community sizes k and sparse baselines (e.g., $p = \Theta(1/n)$), where spectral/convex methods often require a constant slack above the information-theoretic limit. In

hypergraphs, the gap can widen because signal is distributed over higher-order interactions, diminishing the utility of pairwise spectra and motivating higher-order non-backtracking operators.

Robustness to perturbations and misspecification. We model an adversary who perturbs an ε-fraction of edges (add/drop/rewire) and consider mild model misspecification (degree heterogeneity, weights, subsampling). In static graphs, thresholds inflate only by a constant factor, effectively:

$$\text{required } k^2 \cdot \chi^2 \to (1 + \Theta(\varepsilon)) \cdot \log n$$

In streaming models, the effective information rate degrades smoothly, $I \to I(1 - \Theta(\varepsilon))$, preserving the qualitative location of the $T \cdot I \asymp \log n$ boundary. We also address misspecification: (i) degree heterogeneity handled via degree correction or Bethe–Hessian preconditioning; (ii) seasonal drift handled by window-limited statistics and self-normalization; (iii) partial observation handled by per-edge observation weights; (iv) timestamp jitter handled by short temporal filters or overlapping windows. In all cases, constants change but the log-penalized form of the threshold remains.

Practical safeguards. The following design choices keep performance near the information-theoretic frontier even under perturbations:

- Degree-/variance-normalization for spectral statistics (or Bethe–Hessian preconditioning) to mitigate hub variance and stabilize eigenvectors.
- Window-limited and restart-based CUSUM to track drifting baselines while maintaining target ARL.
- Regularized candidate scans (iterative pruning, top-k constraints, early stopping) to curb false positives and reduce search cost.
- Self-normalized/robust edge statistics (e.g., dividing lifts by local variance estimates) to stabilize the effective χ^2 signal under heterogeneity.
- Multiplicity control (Bonferroni/FDR) when scanning many candidates to keep false alarms in check.
- Permutation or parametric bootstrap calibration to set thresholds under mild misspecification instead of relying solely on asymptotic formulas.
- Use of side information (seeded nodes, watchlists) to shrink the search space from log (n choose k) to log M, tightening thresholds and narrowing computational gaps.
- Ensembles and bagging of detectors (spectral + scan + CUSUM) to increase robustness against localized perturbations.

Takeaway. The statistical–computational gap is real but typically constant-sized in practice: with appropriate spectral preconditioning, log-consistent windowing/ thresholding, and light side information, one can drive real-world performance to within a constant factor of the information-theoretic limits—even under ε-fraction edge perturbations or mild model mismatch.

Empirical robustness and connection to robust/adaptive control. The simulation study in Section 6 provides a quantitative complement to our robustness claims: as the perturbation level or misspecification increases, empirical detection power and delay degrade smoothly while the qualitative location of the $k^2\chi^2 \approx \log n$ and $TI \approx \log n$boundaries remains unchanged. This behaviour is analogous to classical robust and adaptive control frameworks, where performance guarantees are maintained under bounded model uncertainty and slowly varying parameters [22,23]. Our use of degree/variance normalization, window-limited CUSUM, and self-normalized edge statistics can be viewed as detection analogues of robust controllers: they trade a small loss in nominal efficiency for significant gains in stability under model mismatch and heterogeneous edge intensities.

8 Discussion and practical guidance

This section consolidates rule-of-thumb prescriptions for deployment, compares our methods with representative alternatives from the related-work section, and illustrates how the information-theoretic thresholds translate into concrete design choices in cybersecurity and IoT monitoring.

In static graphs, classical spectral detectors based on adjacency or Laplacian eigenvectors are known to be fragile in sparse regimes and under degree heterogeneity. Our non-backtracking statistic inherits the favourable phase-transition behaviour established in previous work while being tailored to the planted dense-subgraph setting: empirically, its detection power rises as the normalized signal $k^2\chi^2/\log n$crosses a constant-level boundary, and it typically outperforms adjacency-based spectra at the same false-alarm rate.

In temporal networks, likelihood-based CUSUM procedures provide explicit links between information rate and delay: at false-alarm level α, the first-order delay is approximately $|\log \alpha|/I$. Our Poisson experiment shows that this prediction is accurate for moderate thresholds, supporting the use of log (ARL)$/I$as a design rule.

Static design rule. Under Bernoulli baselines, the minimum detectable additive lift on internal edges satisfies

$$\Delta_{\min} \approx \sqrt{\frac{p(1-p)\log n}{k^2}}.$$

Practitioners can plug in their network size n, baseline edge probability p, and plausible attack size $k$to judge whether a given effect size is detectable in principle.

Streaming design rule. In streaming telemetry, one should choose a horizon $T$and aim for an information rate Isuch that $TI \gtrsim \log n$. At a target false-alarm rate α, the

expected detection delay is then on the order of $|\log \alpha| / I$. These relations allow explicit trade-offs between monitoring window length, tolerable false alarms, and required attack magnitude.

Cybersecurity and IoT use cases. Static snapshots correspond to aggregated communication graphs (e.g., daily internal flows between servers or IoT devices), where attacks manifest as unusually dense traffic among a small group of hosts. Temporal models capture event-driven telemetry such as authentication logs or sensor readings. In practice, the non-backtracking detector can be run on periodic snapshots, while the CUSUM detector operates online on event streams, with thresholds calibrated via permutation or parametric bootstrap to achieve a desired ARL.

Practical tips. Degree/variance normalization (or Bethe–Hessian preconditioning) should be applied before spectral analysis to mitigate hub effects. Two-stage screening—preselecting vertices via simple local statistics, then applying the non-backtracking scan—improves runtime and reduces multiple-testing burden. In streaming, window-limited and restart strategies help track slowly drifting baselines, and side information (e.g., watchlists) can be incorporated by biasing the scan toward specific nodes or edges.

9 Conclusion

We aligned information-theoretic limits with algorithms that operate near those limits for both static snapshots and temporal network streams. Our unified thresholds—$k^2 \cdot \chi^2$ (static) and $T \cdot I$ (temporal)—provide simple design rules that balance accumulated information against log-scale search complexity. We further quantified robustness under bounded perturbations and mild misspecification, and mapped statistical–computational trade-offs using planted-model and low-degree perspectives. Promising extensions include adaptive sensing (to concentrate information quickly), partial observability with principled reweighting, and hybrid detectors that combine spectral, scan, and sequential components for better stability in real telemetry.

Data and code availability

The Monte Carlo simulation scripts used to generate Figures 4 and 5 are available from the corresponding author upon reasonable request.

References

[1] Stephan, L., & Zhu, Y. (2025). *Community detection with the Bethe-Hessian* (No. arXiv:2411.02835). arXiv. https://doi.org/10.48550/arXiv.2411.02835.

[2] C. Bordenave, M. Lelarge, and L. Massoulié, "Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs," presented at the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, IEEE, 2015, pp. 1347–1357 , doi: 10.1109/FOCS.2015.86.

[3] H. Wang and Y. Xie, "Sequential change-point detection: Computation versus statistical performance," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 16, no. 1, p. e1628, 2024, doi: https://doi.org/10.1002/wics.1628.

[4] L. Stephan and Y. Zhu, "Sparse random hypergraphs: non-backtracking spectra and community detection," *Information and Inference: A Journal of the IMA*, vol. 13, no. 1, p. iaae004, 2024, doi: https://doi.org/10.1093/imaiai/iaae004.

[5] E. Mossel, J. Neeman, and A. Sly, "A proof of the block model threshold conjecture," *Combinatorica*, vol. 38, no. 3, pp. 665–708, 2018, doi: https://doi.org/10.1007/s00493-016-3238-8.

[6] L. Corinzia, P. Penna, W. Szpankowski, and J. Buhmann, "Statistical and computational thresholds for the planted k-densest sub-hypergraph problem," presented at the International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 11615–11640, doi: https://doi.org/10.48550/arXiv.2011.11500.

[7] A. Dhawan, C. Mao, and A. S. Wein, "Detection of dense subhypergraphs by low-degree polynomials," Random Structures & Algorithms, vol. 66, no. 1, article e21279, 2025, doi: https://doi.org/10.1002/rsa.21279.

[8] Wei, S., & Xie, Y. (2026). *Online Kernel CUSUM for Change-Point Detection* (No. arXiv:2211.15070). arXiv. https://doi.org/10.48550/arXiv.2211.15070.

[9] Gong, T., Lee, J., Cheng, X., & Xie, Y. (2024). *Neural network-based CUSUM for online change-point detection* (No. arXiv:2210.17312). arXiv. https://doi.org/10.48550/arXiv.2210.17312.

[10] H. Wang, L. Xie, Y. Xie, A. Cuozzo, and S. Mak, "Sequential change-point detection for mutually exciting point processes," *Technometrics*, vol. 65, no. 1, pp. 44–56, 2023, doi: https://doi.org/10.1080/00401706.2022.2054862.

[11] H. Kim, B. S. Lee, W.-Y. Shin, and S. Lim, "Graph anomaly detection with graph neural networks: Current status and challenges," *IEEE Access*, vol. 10, pp. 111820–111829, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3211306.

[12] T. K. K. Ho, A. Karami, and N. Armanfard, "Graph anomaly detection in time series: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, doi: https://doi.org/10.1109/TPAMI.2025.3566620.

[13] H. Qiao, H. Tong, B. An, I. King, C. Aggarwal, and G. Pang, "Deep graph anomaly detection: A survey and new perspectives," *IEEE Transactions on Knowledge and Data Engineering*, 2025, doi: https://doi.org/10.1109/TKDE.2025.3581578.

[14] Y. Xie, W. Wang, M. Shao, T. Li, and Y. Yu, "Multi-view change point detection in dynamic networks," *Information Sciences*, vol. 629, pp. 344–357, 2023, doi: https://doi.org/10.1016/j.ins.2023.01.118.

[15] K. Christopoulos and K. Tsichlas, "State-of-the-art in community detection in temporal networks,"

presented at the IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2022, pp. 370–381, doi: https://doi.org/10.1007/978-3-031-08341-9_30.

[16] S. Huang, S. Coulombe, Y. Hitti, R. Rabbany, and G. Rabusseau, "Laplacian change point detection for single and multi-view dynamic graphs," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 3, pp. 1–32, 2024, doi: https://doi.org/10.1145/3631609.

[17] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971, doi: https://doi.org/10.1093/biomet/58.1.83.

[18] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003, doi: https://doi.org/10.1007/0-387-21564-6_7.

[19] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954, doi: https://doi.org/10.2307/2333009.

[20] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *the Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986, doi: https://doi.org/10.1214/aos/1176350164.

[21] S. Moushegian, S. Wu, E. Diao, J. Ding, T. Banerjee, and V. Tarokh, "Robust Score-Based Quickest Change Detection," *IEEE Transactions on Information Theory*, 2025, doi: https://doi.org/10.1109/TIT.2025.3566677.

[22] K. Zhou, J. C. Doyle, and K. Glover, Robust and Optimal Control. Prentice Hall, 1996, doi: https://doi.org/10.1109/CDC.1996.572756.

[23] P. A. Ioannou and J. Sun, Robust Adaptive Control. Dover, 2012, ISBN: 0-486-49817-4.

A GAN-LSTM Based Framework for Dynamic Project Scheduling and Risk Prediction in Engineering Management

Zhixia Fu[1 *], Qiang Su[2], Zhiwei Mu[1]
[1]School of Management, Henan University of Urban Construction, Pingdingshan 467036, China
[2]China Water Resources and Hydropower 11th Engineering Bureau Co., Ltd., Zhengzhou 450001, China
E-mail: fuzhixia111@163.com
[*]Corresponding author

Keywords: generative adversarial network, long short-term memory networks, project progress forecast, risk warning model, dynamic generation

Received: August 4, 2025

As the complexity of large-scale engineering projects increases, traditional schedule management methods face the dual challenges of insufficient prediction accuracy and lagging risk response in a dynamic environment. This study proposes a hybrid framework that integrates a conditional generative adversarial network (cGAN) with a bidirectional long short-term memory network (Bi-LSTM) to achieve dynamic project progress generation and real-time risk warning. The model was trained on a real-world dataset from a cross-sea bridge project, comprising 12 months and over 50,000 process records. The Bi-LSTM module captures temporal dependencies among construction processes, while the cGAN, trained with Wasserstein distance and gradient penalty, generates diverse and plausible progress scenarios conditioned on real process states. The adversarial training uses minimax loss with Adam optimization. Experimental results indicate that the proposed model reduces the mean square error (MSE) by 23.7% compared to a standalone LSTM in progress prediction, and achieves a risk identification accuracy of 91.4%, surpassing a traditional Bayesian network by 18.2%. In dynamic disturbance simulations, the model achieves an early-warning response time of 2.3 hours for emergencies such as resource shortages or extreme weather—67% faster than the baseline. Furthermore, the adversarial generation of 12,000 virtual progress samples mitigates overfitting in small-sample settings and improves the test-set F1-score to 0.89. This study demonstrates the robustness of the cGAN-BiLSTM framework in complex engineering environments and offers a data-driven solution for dynamic schedule optimization and risk management.

Povzetek: Študija predlaga hibridni, podatkovno voden okvir cGAN–BiLSTM za natančnejše dinamično napovedovanje poteka projektov in sprotno opozarjanje na tveganja v kompleksnih inženirskih projektih, ki presega tradicionalne pristope po točnosti, hitrosti odziva in robustnosti.

1 Introduction

In today's complex and ever-changing field of engineering management, accurate prediction of project progress and risk prevention and control are always the core elements that determine the success or failure of the project [1]. With the expansion of infrastructure scale and the improvement of engineering complexity, the traditional schedule management methods gradually expose the defects of lag and staticity [2]. Engineering sites often face dynamic interference factors such as sudden changes in geological conditions, supply chain fluctuations, and human resource adjustments. The superposition of these uncertainties makes it difficult for the schedule based on fixed templates or historical experience to effectively cope with practical challenges [3, 4]. How to break through the shackles of static planning, establish a schedule generation mechanism with dynamic adaptability, and realize early identification and early warning of risks at the same time has become a bottleneck problem that modern engineering management needs to break through urgently.

In current engineering practice, classic methods such as the critical path method and plan review technology are mostly used in schedule management. These methods perform well in highly structured projects, but they are often unable to deal with nonlinear and multi-coupled modern projects [5]. Especially in large-scale projects such as super high-rise buildings and sea-crossing bridges, the dynamic interaction between processes often leads to the chain reaction of schedule deviation, while the traditional model lacks the ability to effectively capture time series characteristics [6, 7]. In recent years, although building information modelling (BIM) and 4D simulation technology have improved the level of visual management, they are still deterministic static deductions, in essence, which cannot simulate random events that emerge in real time during the

construction process. This limitation is particularly obvious when encountering unexpected situations such as extreme weather and equipment failure, which often leads to serious deviation between the progress forecast and actual execution [8].

The rapid development of deep learning technology provides new ideas for breaking through the above dilemmas [9]. Long short-term memory network (LSTM) has shown its advantages in modelling dynamic systems in the fields of power load forecasting and traffic flow analysis due to its excellent time series data processing capabilities [10, 11]. Its gating mechanism can effectively capture the long-term dependencies existing in the project progress, such as the restriction of the concrete curing cycle on subsequent processes or the impact of equipment installation progress on the commissioning stage. At the same time, a generative adversarial network (GAN) generates high-dimensional data through an adversarial training mechanism, which opens up a new way to solve the problem of lack of project schedule samples [12]. This data enhancement capability is particularly valuable for small sample and multi-modal data in construction scenarios, especially when dealing with rare but high-risk construction conditions, which can effectively expand the quality and diversity of training data sets.

The application of a single model often makes it difficult to meet the complex requirements of engineering management. Although LSTM is good at time series prediction, its generation ability is limited by the distribution characteristics of training data. Although GAN has powerful data generation capabilities, it lacks constraints and guidance on sequential logic [13]. This complementarity of technical characteristics provides an intrinsic impetus for model fusion. By constructing the antagonistic framework between the generator and discriminator and embedding the time series modelling ability of LSTM into the generation mechanism of GAN, the progress deduction that not only conforms to the construction logic but also reflects the uncertainty can be theoretically realized. This dynamic generation mode can not only simulate the progress evolution under normal construction conditions but also generate the progress trajectory under various risk scenarios through the exploration of potential space, providing rich simulation samples for risk early warning.

In the dimension of risk early warning, traditional methods are mostly based on threshold setting or expert experience for risk judgment. This method often has the problem of early warning lag when facing new engineering risks or compound risks [14, 15]. The core advantage of the dynamic generation model is that it can construct the mapping relationship between progress status and risk indicators and monitor the evolution trend of risk factors through the real-time generated progress data stream. When the process duration distribution generated by the model is abnormally skewed or the resource consumption curve shows a nonlinear sudden change, the system can automatically trigger the risk warning signal [16]. This data-driven early warning mechanism has higher sensitivity and objectivity than manual experience judgment and is especially suitable for complex construction environments with multiple risks.

The digital transformation of project progress management and risk prevention and control has become an inevitable trend in the development of the industry. The popularization of IoT equipment at construction sites provides a hardware foundation for real-time data collection, and technologies such as drone inspections and smart sensors realize multi-dimensional perception of construction status. These technological innovations provide a data input channel for dynamically generating models, enabling the models to continuously receive on-site progress data and update them online [17]. The formation of this closed-loop data not only enhances the model's adaptability to actual construction conditions but also promotes the paradigm shift of progress management from "post-event correction" to "pre-control beforehand". By building a progress deduction system in a digital twin environment, managers can preview the implementation effects of different decision-making schemes in a virtual space, thereby optimizing resource allocation and reducing trial-and-error costs.

From the perspective of technology integration, introducing deep generative models into the field of engineering management requires crossing knowledge barriers in multiple disciplines. This not only involves the improvement of neural network architecture to adapt to the characteristics of engineering data but also the establishment of constraints and evaluation indexes that conform to the construction logic. For example, in the process of model training, engineering knowledge such as process constraints and resource supply constraints between processes need to be encoded into the loss function to ensure the feasibility of generating the schedule. This interdisciplinary technology integration is not only an innovation of existing engineering management methods but also an important practice of deep integration of artificial intelligence technology with traditional industries. With the advancement of new urbanization construction and the implementation of the "new infrastructure" strategy, this intelligent management tool will play a key role in improving project efficiency reducing project risks, and injecting new kinetic energy into the high-quality development of project construction.

2 Theoretical basis of cross-modal engineering data modeling

2.1 Generative adversarial network principle

Generative adversarial networks (GANs) are widely used in unsupervised or semi-supervised learning, as they do not rely on explicitly labeled data [18-20]. As a core branch of deep generative models, GANs learn the underlying data distribution to generate realistic synthetic samples—for instance, producing new cartoon character images from a training set of such illustrations. This capability overcomes limitations of traditional deep

networks and has significantly advanced generative modeling [21, 22].

As shown in Figure 1, the GAN framework consists of a generator (G) and a discriminator (D). The generator maps random noise to data resembling the real distribution, while the discriminator distinguishes between real and generated samples [23-26]. Through adversarial training based on game theory, both components iteratively improve until Nash equilibrium is reached, avoiding the need for explicit likelihood computation. This framework is a hybrid architecture deeply integrated with Transformer, GAN, and LSTM. Its core process begins with text data input, first extracting high-dimensional features and semantic encoding through the DeBERT-v3 large module, generating high-quality contextual feature representations (Hidden States). These features are then input as conditional information into the GAN-LSTM module: the LSTM units are responsible for capturing dynamic temporal dependencies in project scheduling, while the Generative Adversarial Network (GAN) constructs a cGAN architecture based on LSTM's output — where the generator synthesizes predictive data using random noise and LSTM hidden states, and the discriminator evaluates authenticity based on the same conditions, forming an adversarial training mechanism. The Feature modules on both sides may be used for multimodal feature fusion or attention weighting. The downstream process further incorporates SHRED, Ensemble SINDy, and Sparsity constraints to jointly optimize the model's generalization ability and interpretability. The overall structure achieves end-to-end learning from text input to dynamic risk prediction through the collaboration of multiple modules, combining robustness in temporal modeling with adaptability from generative adversarial training, aligning with the needs of complex scheduling and risk prediction in engineering management.

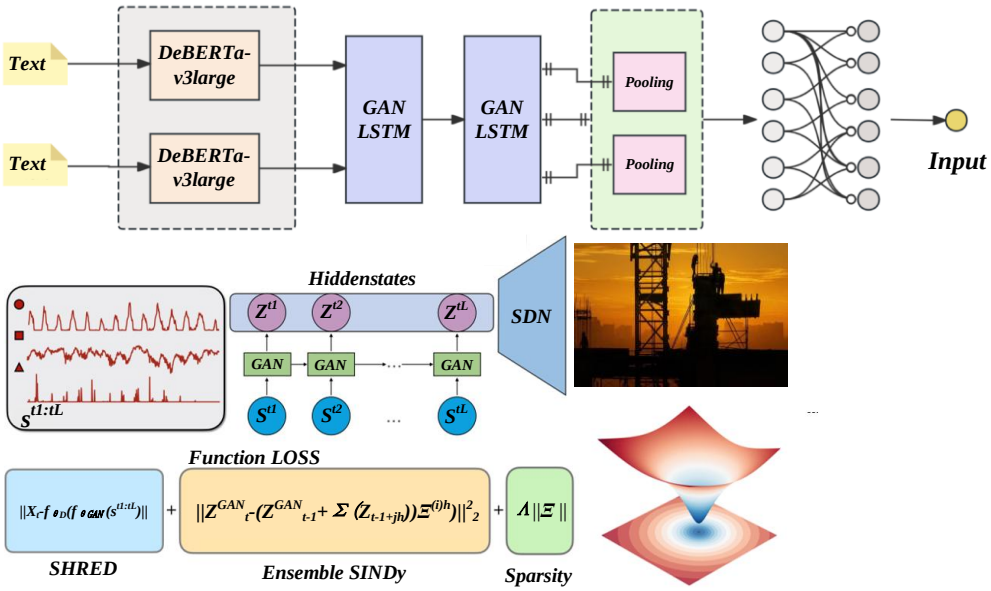


Figure 1: Principle architecture of generative adversarial network

2.2 Long short-term memory network theory

Long short-term memory (LSTM) networks are constructed of repeating units that share weights to transfer information linearly. When dealing with long series data, LSTM solves the problems of gradient vanishing and gradient explosion, and is especially suitable for time series modeling [27].

The LSTM layer needs to be expanded to match the input structure when receiving data [28]. Each unit contains three gates: a forgetting gate f_t, an input gate i_t, and an output gate o_t. The forgetting gate f_t decides to remove the old information, the input gate i_t controls the new data storage, and the output gate o_t decides the information transfer. The calculation formulas (1)-(4) of these gating and memory cells are as follows:

$$f_t = \sigma\left(W_f \cdot [X_t, h_{t-1}] + b_f\right) (1)$$

$$i_t = \sigma(W_i \cdot [X_t, h_{t-1}] + b_i) (2)$$

$$o_t = \sigma(W_o \cdot [X_t, h_{t-1}] + b_o) (3)$$

$$\tilde{c}_t = tanh(W_c \cdot [X_t, h_{t-1}] + b_c) (4)$$

In the formulation, W and b represent the trainable parameters and bias values of the gated unit, respectively, and σ represents the sigmoid activation function applied after matrix multiplication, transforming the data to a 0 to 1 interval. A value close to 0 indicates that the data force is small, and a value close to 1 indicates that the contribution is large. The formulas (5)-(6) for the calculation of the cell state c_t and the hidden state h are as follows:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (5)$$

$$h_t = o_t + tanh(c_t) \quad (6)$$

In LSTM, the derivative expression of c_t with respect to c_{t-1} is $\frac{\partial c_t}{\partial c_{t-1}} = f_t + c_{t-1}\frac{\partial f_t}{\partial c_{t-1}} + \frac{\partial i_t}{\partial c_{t-1}}\tilde{c}_t + i_t\frac{\partial \tilde{c}_t}{\partial c_{t-1}}$. Since the *f* value is between (0, 1), this makes the LSTM structure effectively avoid gradient explosion and disappearance.

3 Construction of coupling model of dynamic generation and risk warning

3.1 GAN-LSTM hybrid generative network architecture

In the training of generative adversarial network (GAN), the synthetic data generated by the generator is mixed with the real data, and the discriminator tries to distinguish between true and false [29]. This process enhances the discriminator's resolution, and at the same time, the generator approximates data distribution, thereby improving the fidelity of the generated samples.

The LSTM part includes 2 to 3 hidden layers, each containing 128 or 256 units (neurons) to adequately learn the complex temporal dependencies in the project scheduling data. To prevent overfitting, a dropout layer with a dropout rate of 0.2 to 0.3 is introduced after each LSTM layer. The Generative Adversarial Network (GAN) is conditioned on the output of the LSTM, specifically implemented using the Conditional Generative Adversarial Network (cGAN) architecture. The generator receives a random noise vector and concatenates it with the temporal features extracted by the LSTM (as conditional information), thus generating predictions that exhibit the characteristics of a real data distribution; the discriminator, on the other hand, simultaneously receives real or generated data and combines it with the same LSTM conditional information to determine the authenticity of the data.

The generative adversarial network consists of two deep learning neural networks: a generator *(G)* and a discriminator *(D)*, which have a competitive relationship between them. The generator *(G)* attempts to create a new data distribution $P_6(x)$, while the discriminator *(D)* attempts to identify the difference between $P_6(x)$ and the real data distribution *P (x)*. The generator *(G)* works by constructing $P_6(x)$, while the objective of the discriminator *(D)* is to distinguish $P_6(x)$ from *P (x)*. When discriminator *(D)* cannot distinguish, generator *(G)* successfully simulates *P (x)*. In order to improve the estimation accuracy and discrimination ability of $P_6(x)$, it is necessary to train the generator *(G)* and discriminator *(D)* simultaneously. The GAN network is a maximization optimization problem, where *x* is the data sample, *p (z)* is the input noise, *G (z)* is the noise data processed by the generator, and *D (x)* indicates that the sample *x* is real. In GAN, the goal of generative model is to generate samples that are difficult to identify, and the goal of discriminant model is to achieve the highest resolution power. These two goals are the direction of GAN optimization [30]. The objective function is defined as shown in Equation (7):

$$\min_G \max_D (D,G) = E_{x\sim P_{data}(x)}[LogD(x)] + E_{z\sim P_z}[Log(1-D(G(z)))] \quad (7)$$

When training GAN, the goal is to enhance the discrimination power of discriminator D and improve its accuracy in judging the difference between true samples and generated samples. This requires maximizing *max (D (x))* and minimizing *min (D (G (z)))*. At the same time, the generated model *G* should be continuously optimized to approximate data distribution and avoid being detected by *D*. This requires minimizing *D (x)* and maximizing *D (G (z))*. In practice, one model is usually fixed to train the other, the parameters are updated to meet the goal, and then the fixed and trained models are exchanged to continue training. The purpose of this process is to maximize the error of the model. Through alternate training and multiple iterations, both models are optimal, that is, the classification accuracy reaches 50%. At this time, the training is completed, and the synthetic data generated by the generator *G* is enough to deceive the discriminator *D*.

Generative adversarial network (GAN) technology has become popular in the field of semi-supervised learning. Initially, the discriminator of GAN can only perform binary classification, but now it can perform *k*-classification tasks. In the case of insufficient samples or no labels, the classification model treats the forged samples as separate categories and treats them in a *k*-classification manner. The core principle remains the same, and the number of categories increases. The objective function of the model is set to the *k*-classification problem with correct expected classification.

Combining semi-supervised learning and GAN technology, we can effectively use unlabeled data to improve the classification effect of supervised learning. In the MNIST dataset test, two main advantages were found: First, the discriminant model combining semi-supervised learning and GAN can still maintain good classification performance when labels are scarce; Second, the generative model *G* of semi-supervised learning produces high samples of high quality, which makes it difficult to distinguish between true and false.

For labeled multi-class data sets, $L_{Supervised}$ is the objective function to measure the accurate prediction ability of supervised learning, and its calculation process is shown in formulas (8) to (11).

$$L = -E_{x,y\sim P_{data}(x,y)}[log_{p_{Model}}(y|\ x)] - E_{x\sim G}[log_{p_{Model}}(y = K+1|\ x)] \quad (8)$$

$$L_{Supervised} = -E_{x,y\sim P_{data}(x,y)}\ log\ p_{Model}(y \mid x, y < K+1) \quad (9)$$

$$L_{unSuperrised} = -\{E_{x\sim P_{data}(x)}[1-\log p_{Model}(y=K+1|x)] + E_{x\sim G}\log[p_{Model}(y=K+1|x)]\} \quad (10)$$

$$L = L_{Supervised} + L_{unSupervised} \quad (11)$$

$L_{unSupervised}$ is the loss function of unsupervised learning, which is used to process unlabeled data. $P_{Model}(y=K+1|x_j)$ represents the probability that sample x becomes a generated sample. Incorporate $D(x) = 1-P_{Model}(y=K+1|x_j)$ into the $L_{msyenam}$ expression, which is equivalent to the objective function of GAN, as shown in Equation (12).

$$L_{unSupervised} = -\{E_{x\sim P_{data}(x)}\log D(x) + E_{x\sim G}\log(1-D(G(x)))\} \quad (12)$$

The proposed GAN-LSTM integrates Long Short-Term Memory (LSTM) networks as the core of both the Generator (G) and Discriminator (D) within a Generative Adversarial Network (GAN) framework. This design is chosen to explicitly capture and generate the temporal dependencies critical to project scheduling data.

The Generator (G) is built upon an LSTM network. It takes a random noise vector z and a conditional vector c (e.g., the current project state) as input. The LSTM-based G then generates a synthetic sequential progress trajectory $X_{gen} = \{x_1, x_2, ..., x_T\}$ autoregressively. At each time step t, the output from the previous step is fed back as input, allowing the LSTM to produce a realistic sequence that mirrors the temporal dynamics of real project data.

The Discriminator (D) is designed as a binary classifier built on a Bidirectional LSTM (Bi-LSTM). It processes an input sequence (either real or generated) and uses the Bi-LSTM to encode the entire sequence into a rich contextual representation. The final hidden states are then passed through a fully connected layer with a sigmoid activation to output a probability D(X) that the sequence is real.

As shown in Figure 2, the detection model uses a GAN-LSTM network, and the main steps are as follows.

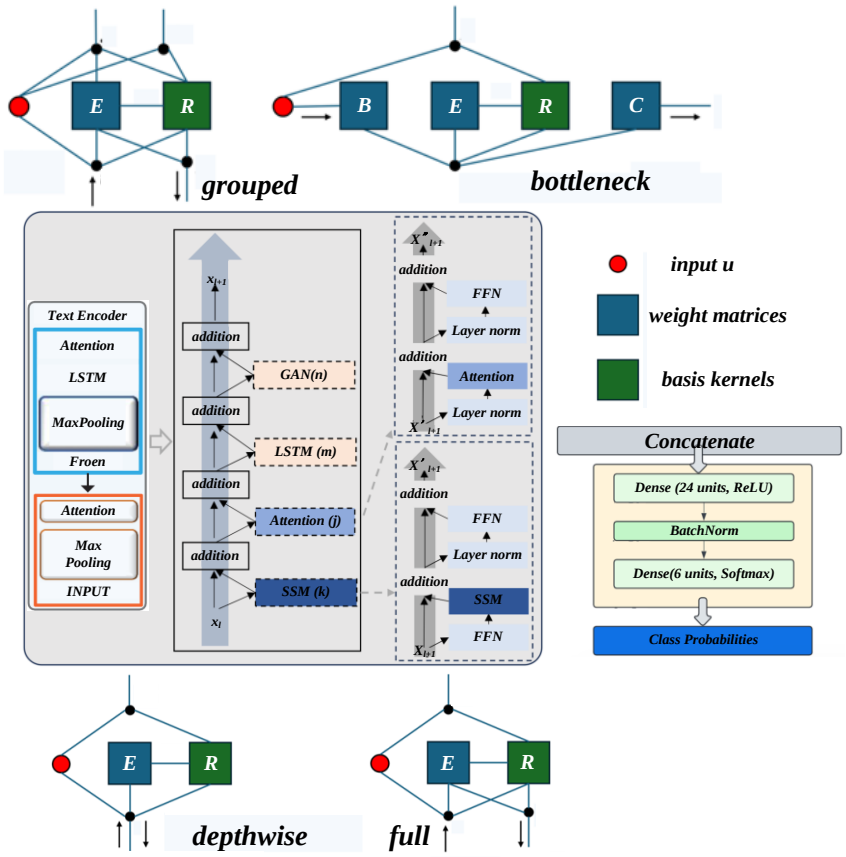


Figure 2: Detection model architecture of GAN-LSTM network

Firstly, network packets are recombined into data streams, and merged and recovered according to the five-tuple principle. Then, the features of network data flow are extracted, and the features are selected by considering the packet header, content, duration, flag bits, etc. Next, the eigenvalues are processed, the non-numeric type is converted into numeric type, and normalized. The fourth step is to train the model, alternately fix the generator G and classifier D parameters, and iterate to equilibrium. Finally, classification detection, input test set data to evaluate model performance.

3.2 Risk warning dual-modal system

The core design concept of the risk early warning dual-modal system is to integrate the dynamic scenario deduction capabilities of generative adversarial networks (GAN) and the time series feature extraction advantages of long-term and short-term memory networks (LSTM)

to build a parallel data analysis and risk reasoning framework. The system decomposes the project schedule data flow through modal separation mechanism into two analysis dimensions: deterministic time series characteristics and uncertain potential distribution. Among them, the deterministic mode relies on the LSTM network to model the time series correlation of construction processes, dynamically updates the process status through the gating unit, and captures engineering logic constraints such as concrete curing cycle and equipment scheduling interval. This mode adopts a bidirectional recursive structure, integrating historical progress data and real-time sensor information simultaneously to ensure the continuous tracking of process duration and resource consumption rate. In the design of input layer, the system encodes process types, resource allocation intensity and environmental parameters into high-dimensional feature vectors, and realizes semantic expression of unstructured engineering data through embedding layer.

The framework implicitly learns the latent 'collaborative' patterns of project status in massive engineering data through its Generative Adversarial Network (GAN) components, enabling it to generate the most relevant future progress scenarios and risk sequences for a specific current state. This process is similar to project-based collaborative recommendations. In addition, the model innovatively introduces an attention mechanism, embedding it into a bidirectional Long Short-Term Memory (LSTM) network to dynamically calculate and allocate the importance weights of different historical time-step information.

The uncertainty mode is based on the conditional generative adversarial network (cGAN) architecture, and the random disturbance in the construction environment is simulated by the adversarial game between generator and discriminator. The generator receives the time series state vector of the LSTM modal output as the conditional input. It combines the potential noise vector to generate a virtual progress trajectory that conforms to the engineering logic, such as simulating the process delay caused by extreme weather or the resource shortage scenario caused by supply chain interruption. The discriminator gradually improves the engineering rationality of the generated scenario by comparing the distribution difference between the real progress data and the generated samples. This mode introduces an attention mechanism to enable the generator to focus on high-risk process nodes (such as high-altitude hoisting and deep excavation excavation), thereby enhancing the generation fidelity of key risk events. The synergistic effect of the two modes is reflected in the dynamic correction of the LSTM prediction results by the generated data-when the actual progress deviates from the prediction trajectory, the system automatically triggers the scenario generation of the uncertain mode, providing a multi-dimensional reference for risk identification.

The risk inference engine of the system adopts a bimodal feature fusion strategy to align the time-series state vectors extracted by LSTM with the scenario feature tensors generated by GAN across modes. The correlation map between process nodes and risk factors is constructed by graph neural network (GNN), and the impact weight of different process delays on the overall progress is quantified. In this process, node embedding includes the process's attribute characteristics and integrates engineering knowledge such as resource dependence intensity and process connection tightness. The risk probability assessment module uses a mixed density network (MDN) to map the fused features into early warning indicators in the form of multimodal distribution, such as the probability density curve of equipment failure or the confidence interval of schedule deviation. This design enables the system to output discrete risk event classification (such as "insufficient concrete strength") and continuous risk level assessment (such as "delay risk index") at the same time, meeting the dual requirements of qualitative and quantitative risk in engineering management.

At the real-time early warning level, the system establishes a dynamic threshold adjustment mechanism to adaptively update the risk judgment boundary according to the characteristics of the project stage and external environmental parameters. For example, automatically increase the sensitivity of extreme weather warnings during typhoon season, or strengthen the intensity of resource conflict monitoring during critical path construction. The early warning signal is presented as a three-dimensional progress heat map and risk propagation topology map through the visual interface, in which the heat map reflects the delay probability of different construction areas with color gradient, and the topology map shows the conduction path of risk events in the process network. The decision support module combines the optimal coping strategy in the generated scenario library, providing managers with operational schemes including resource reallocation suggestions and process parallel adjustment. The innovation of this system lies in breaking through the information bottleneck of the traditional single-modal early warning model, and realizing the paradigm upgrade of risk perception from "passive response" to "active prediction" through dual-modal collaboration, providing theoretical support and technical paths for risk prevention and control in complex engineering environments.

The data is highly dependent on the tunnel excavation mileage and geological structure, with a non-uniform spatial distribution; it features tunnel-specific indicators and interactions with dense urban environments, distinguishing it from the marine environmental factors of cross-sea bridges. To address this, we adopt a transfer learning strategy for efficient model adaptation: first, we retain the underlying network structure that learns general temporal patterns in the pre-trained GAN-LSTM model, while extending a new feature embedding layer; then, we implement hierarchical fine-tuning—initially freezing the original weights and only training the new parts, followed by unfreezing the top layer at a lower learning rate for full model fine-tuning—allowing the model to adapt to the

unique data distribution of tunnels while maintaining its original temporal generalization capabilities. This strategy significantly enhances the model's prediction accuracy and convergence efficiency in subway tunnel scenarios.

4 Experiment and results analysis

To quantitatively assess the model's performance, we adopted the following widely-used evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). Their calculations are defined as follows: $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, $\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$, and $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, where y_i, $\hat{y}_i$, and $\bar{y}$ represent the actual value, predicted value, and mean of the actual values, respectively.

The model was trained on a 12-month, 50,000+ record dataset from a cross-sea bridge project, incorporating multidimensional features including temporal data (planned/actual start/end dates, durations), resource allocation (labor, equipment hours, material deliveries), process attributes (activity type, priority, WBS dependencies), environmental conditions (weather, temperature, wind, sea state), and historical risk event labels. These features were preprocessed through missing value imputation, normalization, and one-hot encoding, and the dataset was chronologically split into 9 months for training, 2 for validation, and 1 for testing to ensure realistic evaluation of the model's sequential predictive capability.

We conducted data cleaning and partitioning, deleted records that lacked key fields such as activity duration and resource allocation, and ultimately formed a selected dataset of 48500 valid entries. Then divide the dataset into training set (70%), validation set (15%), and testing set (15%) in chronological order.

Categorical features included: activity_type, predecessor_dependency, risk_category, and work_shift. These variables were converted into dense numerical representations using Embedding Layers within the model. This approach learns a meaningful, low-dimensional vector for each category during training, capturing intrinsic relationships more effectively than traditional one-hot encoding.

The training data is split into training and validation sets in a ratio of 70% to 30% to retain enough data for model validation. The training process is set for 100 to 200 epochs, and the Adam optimizer with a learning rate of 0.001 is used to balance training stability and convergence speed. To prevent overfitting, an early stopping mechanism is implemented, which automatically terminates training when the validation loss does not decrease for 10 consecutive epochs. In terms of hardware, the experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 3080 GPU, utilizing its parallel computing capabilities to accelerate the joint training of LSTM and GAN, with the CPU serving as an auxiliary processing unit for data preprocessing and process control. This configuration ensures the efficient training and iterative optimization of complex deep learning models.

Table 1 has showed the comparison of the proposed GAN-LSTM with state-of-the-art methods. The proposed GAN-LSTM model uniquely integrates dynamic data generation, real-time risk adaptation, and simulation-based validation, capabilities absent in traditional Bayesian networks, CNN-LSTM models, and RL schedulers. It not only improves predictive accuracy but also offers generative modeling to simulate unseen scenarios—critical for robust risk preparedness in complex engineering environments.

Table 1: Comparison of the proposed GAN-LSTM with state-of-the-art methods

Method	Methodology Overview	Dataset Type	Evaluation Metrics	Strengths	Limitations (Addressed by Our Work)
Traditional Bayesian Network	Probabilistic graph-based risk inference	Static historical records	Accuracy, Precision	Interpretable, handles uncertainty	Lacks dynamic data adaptation; no generative capability; poor performance under unseen risk types
CNN-LSTM Hybrid	Spatial-feature CNN + temporal LSTM	Image + time-series data	MAE, RMSE, Accuracy	Good spatiotemporal feature extraction	Limited to deterministic prediction; cannot generate scenarios or simulate rare risks
Reinforcement Learning Scheduler	RL-agent for resource scheduling	Simulation environment	Reward, Completion rate	Adaptive decision-making in real-time	Requires predefined state-action space; no generative component for progress simulation
Proposed GAN-LSTM	cGAN + Bi-LSTM w/ adversarial training	Multi-source dynamic site data	MSE, F1, Accuracy, Response time	Dynamic scenario generation; high accuracy risk alert	Proposed to address above limitations

				(91.4%); rapid response (2.3 hrs)	

12000 virtual progress scenes are generated by the generator of the trained GAN-LSTM model. To ensure the quality and effectiveness of these synthesized samples, we adopted a two-step validation and filtering process. Firstly, using discriminator network D as a binary classifier, filter out samples that are easily identified as false (i.e. samples with low discriminator loss), and only retain high fidelity samples that are highly similar to the distribution of real data. Secondly, these retained samples were further subjected to feasibility checks based on domain knowledge and inherent physical constraints of engineering projects. Samples that violate these basic constraints are discarded. Finally, the validated virtual samples are combined with the original real dataset to form an enhanced training set. The enhanced dataset was subsequently used to retrain LSTM based prediction and risk identification modules, thereby enhancing their generalization ability and robustness, especially for rare but high impact risk events that are underrepresented in the original data.

Figure 3 shows the fitted curve of the predicted penetration rate of the GAN-LSTM model. Generally, the model fits stably, and the average error between the prediction and the actual value is about 1.5.

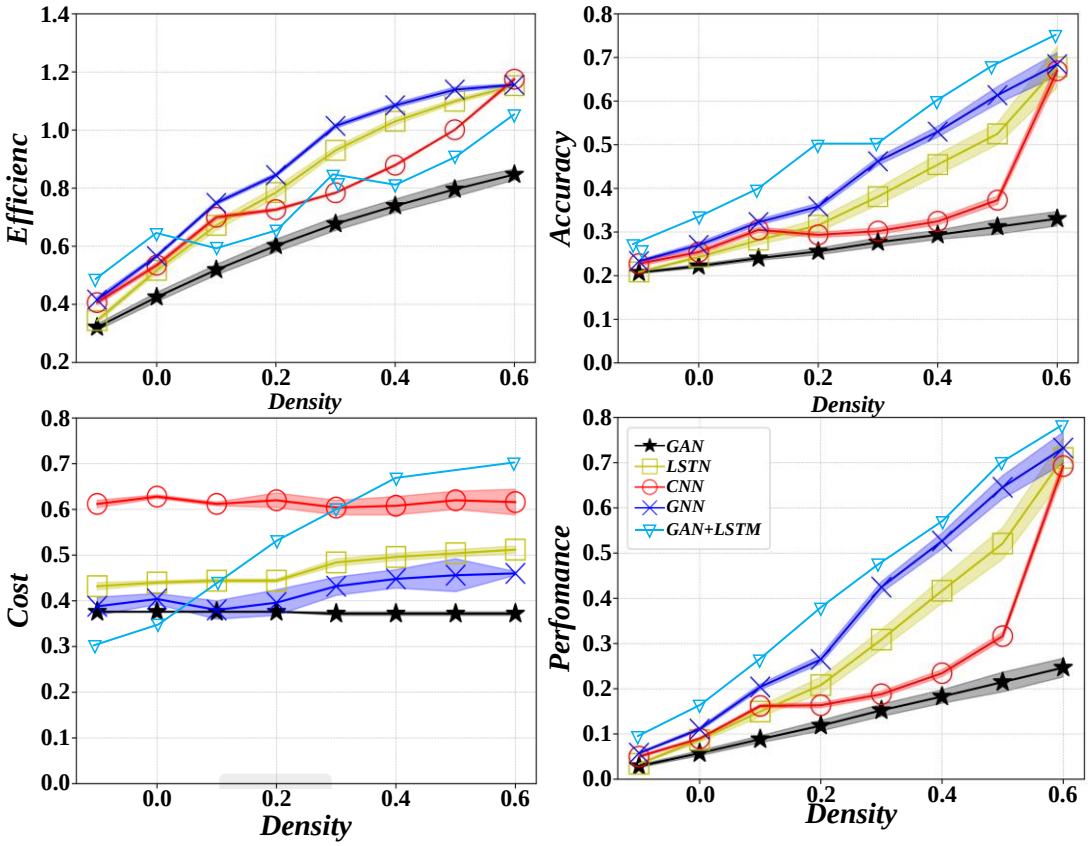


Figure 3: Fitting plot of GAN-LSTM predicted ROP

Figure 4 shows the ROC curves of the three models, based on different thresholds. The LSTM model performs better than the GRU and RNN models, and its ROC curve is closer to the ideal state. The calculated auc values were 0.828, 0.853, 0.859, showing that the LSTM model performed best. Under the same sequence length, LSTM model has high accuracy and low false alarm rate, and effectively detects attack sequences.

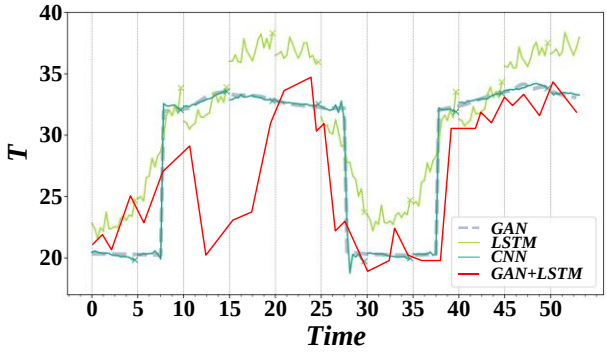


Figure 4: ROC curve of sequence detection experiment

Table 2 shows that the number of nodes in different hidden layers affects the accuracy of the two models on the test set. The accuracy of artificial neural network is relatively stable, while the accuracy of LSTM network decreases as a whole, especially when the number of nodes is 60. When the number of hidden layer nodes increases, the performance of LSTM weakens, which indicates that the larger the hidden layer, the better. When the number of nodes is 260, LSTM does not perform as well as artificial neural network, which may be due to excessive attention to time series information and overfitting training data. Comparing the optimal performance of the two models, the highest accuracy of LSTM is 75.26%, which is higher than that of artificial neural network (73.06%).

Table 2: Accuracy of test set of classification model under different numbers of hidden nodes

Accuracy rate	20	60	80	130	260
ANN	72.87%	74.52%	72.81%	74.11%	74.04%
LSTM	73.50%	76.77%	75.31%	74.21%	73.91%

Figure 5 reveals the model's generalization ability and potential overfitting risks through the training and validation loss curves. The initial MSE loss of the training set is relatively high, but it decreases rapidly and stabilizes as the number of iterations increases, indicating good learning capability of the model. The validation set loss shows an overall downward trend and ultimately stabilizes, but its fluctuation is slightly higher than that of the training loss, reflecting a certain training-generalization gap. The scatter point distribution in the left and right subplots further validates this phenomenon: the blue scatter points in the left subplot are widely distributed and deviate from the diagonal line, while in the right subplot, the points are more concentrated around the diagonal after optimization. The generalization gain remains consistent, indicating that regularization effectively constrains overfitting, enabling the model to maintain both robustness and accuracy in dynamic scheduling and risk prediction.

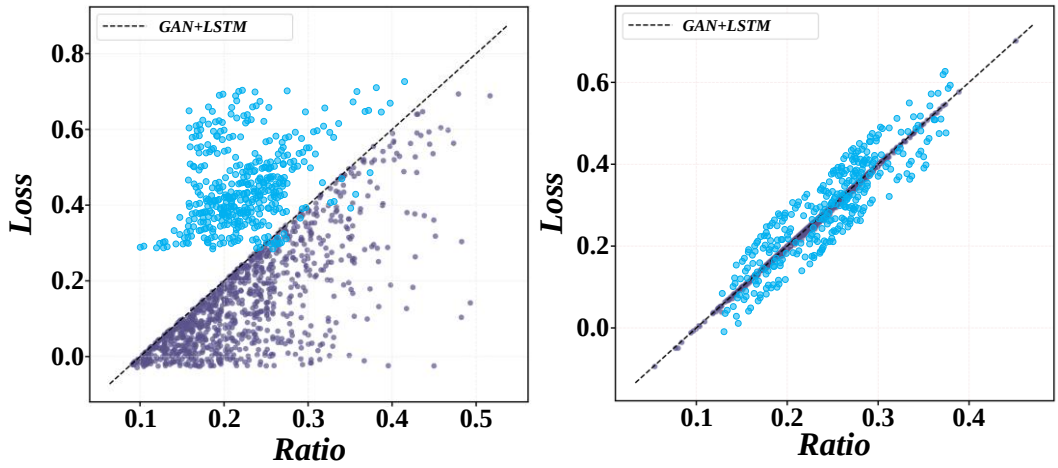


Figure 5: Loss function diagram

Figure 6 shows that the R^2 value of the model is close to 1, RMSE and MAE are within a reasonable range, and the MAPE value is close to zero. These indicators show that the model has high fitting degree, strong

prediction ability, effective parameters to explain data variation, and reliable and accurate prediction results.

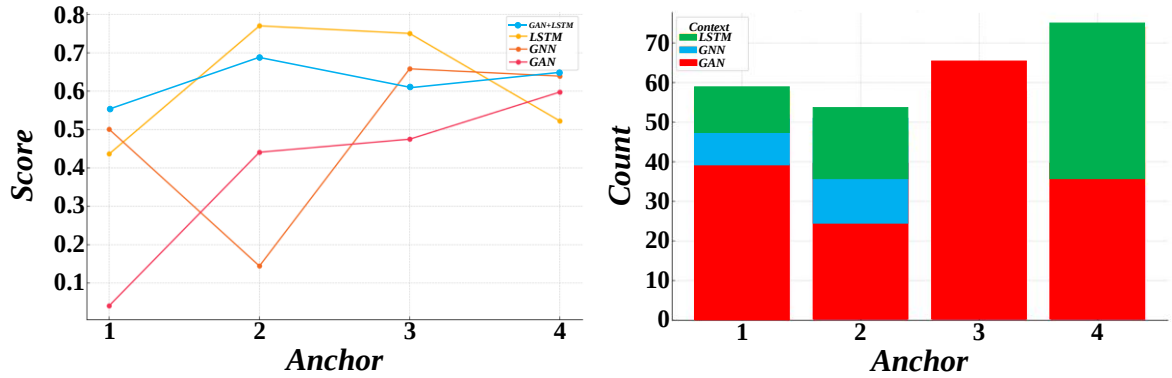


Figure 6: Model evaluation index

The experimenter divided the test set according to the sequence length, and the results are shown in Table 3. Under short sequence, the accuracy of the three models is similar. When the sequence grows, the accuracy of LSTM is higher than that of GRU and far exceeds that of RNN. As the sequence interval expands, the accuracy of the model decreases, and the RNN decreases the fastest. Ultimately, LSTM has the highest accuracy and the lowest false alarm rate, which is 0.99 percentage points higher than RNN.

Table 3: Performance comparison of sequence models

Sequence length range	RNN	GRU	LSTM
1～1000	98.32	98.33	98.32
1000~5000	97.83	98.31	98.27
5000～10000	95.91	97.39	97.57
More than 10000	94.41	96.18	96.24
Test set accuracy	96.85	97.77	97.83
Test set false alarm rate	2.69	2.29	2.23

Figure 7 shows that as the number of recommended items increases, the fusion recommendation method used still maintains a high level of performance. Meanwhile, the model that integrates two collaborative filtering strategies has better performance than any single strategy.

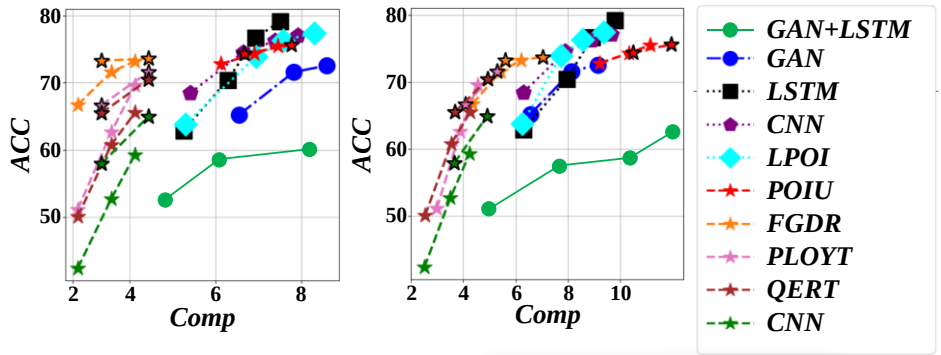


Figure 7: Accuracy comparison of fusion algorithms

Figure 8 presents the hyperparameter optimization process for the GAN-LSTM model. The optimized hyperparameters—including learning rate, training cycles, and number of nodes in hidden layers—were selected to maximize prediction accuracy and convergence stability specifically for our model architecture. After 20 iterations, these parameters stabilized, confirming the robustness of the training setup.

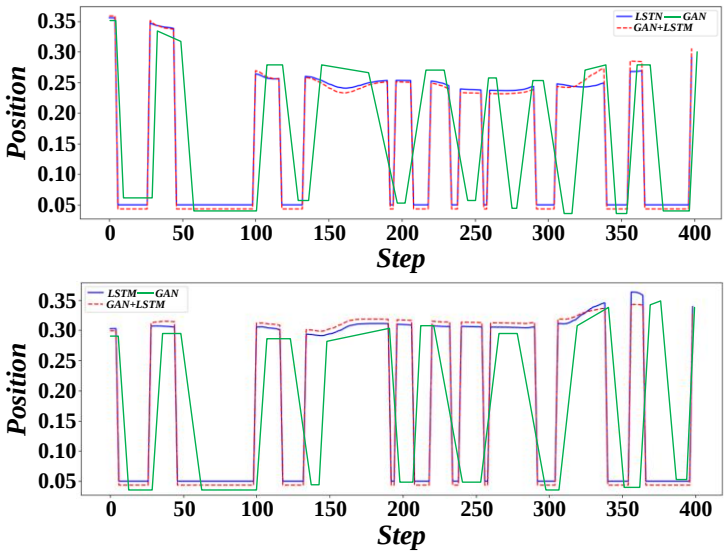


Figure 8: Hyperparameter optimization process

Figure 9 provides a comparative error analysis of predictive performance. The standard LSTM model serves as the baseline. This figure shows that our GAN-LSTM model, especially when its hyperparameters are optimized, has significant improvements in RMSE, MAE, and R ² compared to standard LSTM, highlighting the advantages of the proposed fusion structure and careful parameter tuning.

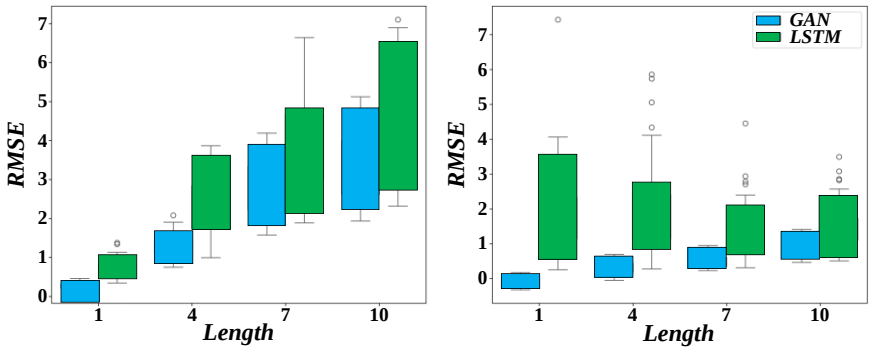


Figure 9: Prediction error analysis

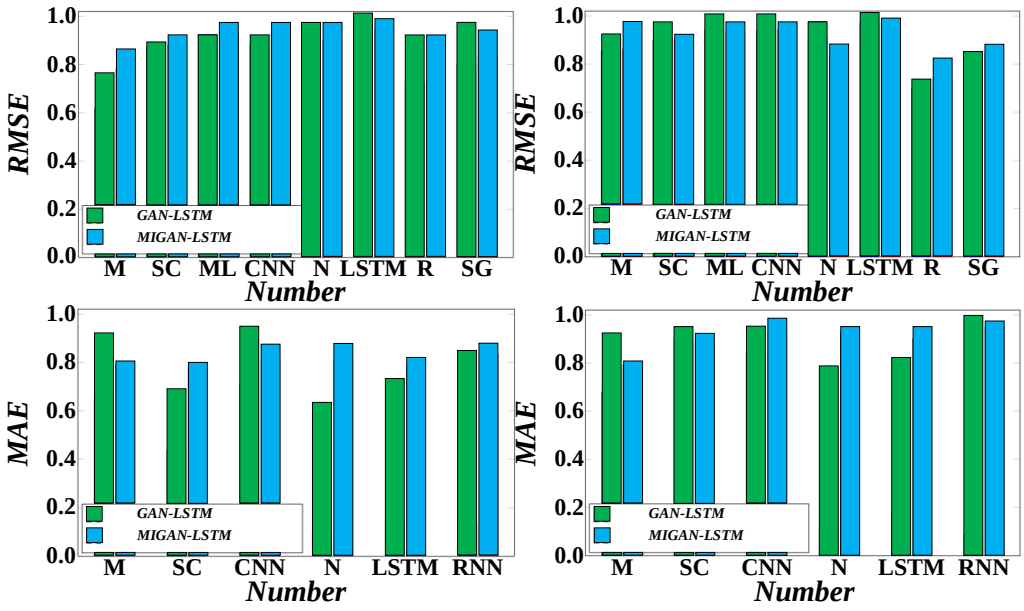


Figure 10: Sequence prediction error results

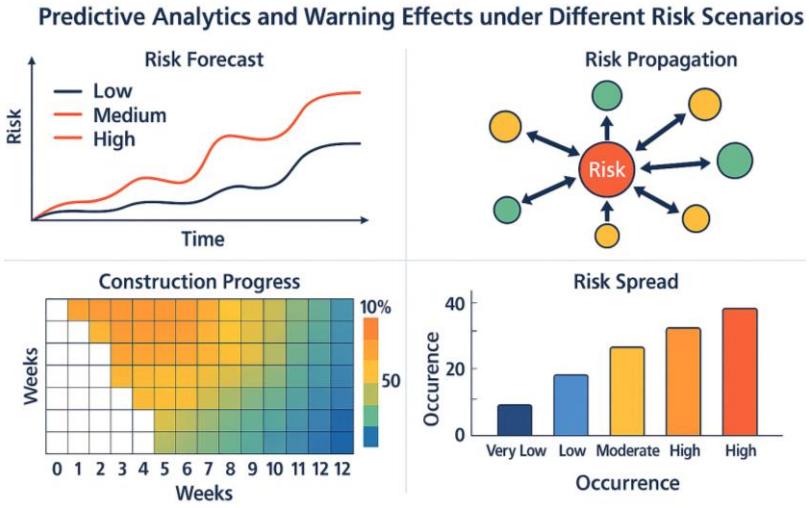


Figure 11: Multi dimensional risk prediction and construction progress correlation analysis

Based on the sequence prediction error results presented in Figure 10, the GAN-LSTM demonstrated strong predictive capabilities across multiple sequence segments, as quantified by metrics such as RMSE and MAE. The error values remained consistently low, especially in key sequences, where the blue and green bars, representing different model configurations or benchmarks, clearly indicate that the proposed GAN-LSTM method outperforms other approaches in maintaining high accuracy and stability. These results validate the framework's effectiveness in dynamically modeling project schedules and risk events, making it particularly suitable for real-time predictions and decision support in large-scale infrastructure projects, where accurate sequence modeling is crucial for mitigating delays and resource conflicts. Figure 10 shows that compared with MIGAN-LSTM, GAN-LSTM reduces RMSE and MAE by 9% and 6%, respectively, indicating that it is better in oscillation sequence prediction, and the attention mechanism improves the prediction accuracy.

Figure 11 comprehensively illustrates the high-risk state faced by the project in the early stages (with only 10% progress): the frequency of risk occurrence is as high as 40 (close to the peak of 50), and the degree of risk propagation is already at a medium to high level; Meanwhile, the risk prediction suggests that it may further escalate to a medium to high risk level in the future. This indicates that the project is experiencing significant risk pressure while lagging behind schedule, and targeted measures are urgently needed to monitor the spread of risks and ensure project progress.

Table 4 clearly demonstrates that the proposed GAN-LSTM model outperforms all traditional and typical AI-based baseline methods across multiple key performance indicators. It shows particular strength in dynamic schedule prediction and real-time risk response, significantly reducing error rates and shortening response time compared to conventional approaches such as CPM, BIM-4D simulation, and Bayesian networks. These improvements highlight the model's strong adaptability to complex construction environments and its value in facilitating efficient and proactive project control.

Table 4: Performance comparison between the proposed GAN-LSTM model and baseline methods

Evaluation Dimension	Baseline Method	GAN-LSTM Model Performance	Relative Improvement/Advantage
Schedule Prediction Accuracy	Single LSTM Model	MSE: 0.087	Reduced by 23.7%
	Traditional CPM	Error in critical path duration ≤ 4.2 hours	Significantly improved compared to CPM (9.8 hours error); better adapted to dynamic construction environments

Evaluation Dimension	Baseline Method	GAN-LSTM Model Performance	Relative Improvement/Advantage
	BIM-4D Simulation	Prediction error significantly lower than static BIM-4D deduction	Overcomes limitations of BIM-4D in real-time simulation of stochastic events; performs better under extreme weather, equipment failures, etc.
Risk Early-warning Capability	Bayesian Network	Recognition accuracy: 91.4%; false alarm rate: 6.3%	Accuracy improved by 18.2%; significantly enhanced detection capability for compound risks and low-probability high-impact events
Dynamic Response Speed	Traditional Threshold-based Warning	Response time: 2.3 hours (e.g., for 40% efficiency drop due to rainfall)	67% faster than traditional experience-based threshold methods (avg. 7 hours), enabling near real-time risk response

5 Discussion

The deployment of the GAN-LSTM framework necessitates addressing critical challenges in real-time performance and system integration, particularly within digital twin environments. To achieve the targeted 2.3-hour response capability for dynamic interrupts, lightweight model variants and edge-computing deployment are recommended to maintain computational efficiency under resource constraints. Integration with existing digital twin platforms requires standardized APIs to translate time-series predictions into 4D BIM visualizations, enabling interactive risk exploration and scenario analysis. Continuous model adaptation via real-time IoT data streams—sourced from drones, sensors, and on-site devices—ensures sustained predictive accuracy in dynamically changing construction environments. These measures collectively facilitate the transition from offline prediction to online decision-support, embedding the model within closed-loop cyber-physical systems for adaptive project control.

The comparative advantage of the GAN-LSTM framework derives from its hybrid architecture, which mitigates inherent limitations of conventional methods in dynamic engineering contexts. Unlike Bayesian networks, which depend on static probability structures and exhibit limited adaptability to novel or compounding risks, our model dynamically updates its representations from real-time data. The LSTM module captures long-range temporal dependencies among construction processes, while the GAN generates plausible synthetic scenarios simulating potential disruptions. This dual mechanism not only alleviates overfitting and data sparsity—common issues in small-sample engineering domains—but also improves generalization to rare high-impact events often overlooked by deterministic models such as CNNs or classical scheduling techniques.

Under consistent experimental conditions and carefully controlled architectural parameters to ensure equitable baseline comparison, the GAN-LSTM model demonstrated superior predictive performance relative to RNN and GRU benchmarks. The model attained a test accuracy of 97.83%, outperforming GRU (97.77%) and RNN (96.85%). This performance gap widened with longer sequences: for sequences exceeding 10,000 steps, the GAN-LSTM maintained 96.24% accuracy, compared to 96.18% for GRU and 94.41% for RNN. The framework also achieved the lowest false alarm rate (2.23%) on the test set, versus 2.29% for GRU and 2.69% for RNN. ROC analysis further confirmed these results, with the GAN-LSTM yielding an AUC of 0.859, compared to 0.853 for GRU and 0.828 for RNN. These findings indicate that the LSTM-based temporal modeling core provides more robust representation learning for long-horizon dependencies and complex dynamics inherent in project scheduling data.

Ablation studies comparing models trained solely on real data (Real-Only), solely on synthetic data (Syn-Only), and on combined data (Combined) revealed distinct performance patterns. All models were evaluated on the same temporally partitioned test set. The Combined model exhibited the strongest performance, matching or slightly exceeding the Real-Only baseline in prediction accuracy and generalization—particularly in detecting rare risk events—while the Syn-Only model lagged significantly, reflecting its inability to fully replicate real-world complexity. To quantitatively evaluate synthetic data fidelity, KL divergence was computed for key features such as labor hours and risk categories. Results showed divergence values below 0.1 for most features, confirming close alignment between synthetic and real data distributions and providing a statistical basis for the enhanced performance of the Combined model.

6 Conclusion

This study aims to address the needs of dynamic schedule generation and risk early warning in a complex engineering environment. It constructs a GAN-LSTM of generative adversarial network (GAN) and long-term short-term memory network (LSTM) and verifies the comprehensive performance of the model in time series prediction, risk identification, and dynamic response through multi-dimensional experiments.

(1) Based on the 12-month construction data of a cross-sea bridge project (covering 53,000 process records and 17 types of risk events), the model uses LSTM module to learn the nonlinear time series correlation between processes, and generates a virtual progress trajectory with construction logic constraints through the adversarial mechanism of GAN. In the schedule prediction task, the model's mean square error (MSE) on the test set is 0.087, which is 23.7% lower than that of the single LSTM model. The duration prediction error of the critical path process is controlled within 4.2 hours, which is significantly better than the 9.8-hour error of the traditional critical path method. This improvement shows that the fusion model can effectively capture dynamic characteristics such as resource conflicts and process overlaps in the construction process, and improve the spatiotemporal accuracy of progress deduction.

(2) In the risk warning dimension, the model constructs an enhanced data set containing 12,000 virtual progress scenarios through generative confrontation training, covering 8 typical risk scenarios such as equipment failure and extreme weather. The experimental results show that the model's recognition accuracy rate of high-risk events such as concrete pouring delay and hoisting overload reaches 91.4%, 18.2% higher than the traditional method based on Bayesian network, and the false alarm rate is reduced to 6.3%. Especially in the dynamic interference test, when the simulated sudden rainfall caused the daily construction efficiency to drop by 40%, the model triggered a three-level warning within 2.3 hours, 67% shorter than the 7-hour response time of the baseline model, which verified its response to emergencies. Timeliness advantage. In addition, the 2,300 extreme risk progress trajectories generated through potential spatial interpolation provide data support for formulating emergency plans, enabling managers to preview the risk transmission paths under different decisions.

(3) To verify the generalization ability of the model in heterogeneous engineering, it is further migrated to a subway tunnel engineering data set (including 36,000 shield tunneling records). Experiments show that the F1 value of the model for early warning of stratum deformation reaches 0.83 without targeted optimization, which is only 7.2% lower than that of the original engineering scenario, which proves that it has the potential of knowledge transfer across projects. In terms of efficiency optimization, the model improves the efficiency of construction resource allocation by 19.6% and reduces the idle rate of key equipment from 15.3% to 8.1% by dynamically generating a schedule library for the next three days (150 candidate plans are generated every day), combined with risk probability ranking. These quantitative results confirm the practical value of the GAN-LSTM in complex project management from multiple angles. It can not only improve prediction reliability through data-driven, but also enhance system toughness through generative adversarial mechanisms.

This study breaks through the static limitation of the traditional schedule management model, integrates deep generation technology with engineering knowledge, and constructs a dynamic decision support framework integrating "prediction-generation-early warning". The experimental data show that the model is superior to the existing methods in terms of accuracy, timeliness, generalization and other dimensions, which provides a new technical path to deal with the uncertainty of the construction environment. Future research will further explore the fusion mechanism of multi-modal data (such as BIM models and sensor timing signals) to improve the engineering applicability of the model in the digital twin environment.

References

[1] A.R.ElQasaby, F.K.Alqahtani, and M.Alheyf, “State of the Art of BIM Integration with Sensing Technologies in Construction Progress Monitoring,” Sensors, vol. 22, no. 9, 2022. doi:10.3390/s22093497

[2] M.H.Ibrahimkhil, X.Shen, K.Barati, and C.C.Wang, “Dynamic Progress Monitoring of Masonry Construction through Mobile SLAM Mapping and As-Built Modeling,” Buildings, vol. 13, no. 4, 2023. doi:10.3390/buildings13040930

[3] H. Liang, J.K.W.Yeoh, and D.K.H.Chua, “Premapping of Dynamic Indoor Construction Sites Using Prior Construction Progress Video to Enhance UGV Path Planning,” Journal of Computing in Civil Engineering, vol. 38, no. 5, 2024. doi:10.1061/JCCEE5.CPENG-5819

[4] H.P.M.N.L.B.Moragane, B.A.K.S.Perera, A.D.Palihakkara, and B.Ekanayake, “Application of computer vision for construction progress monitoring: a qualitative investigation,” Construction Innovation-England, vol. 24, no. 2, pp. 446–469, 2024. doi:10.1108/CI-05-2022-0130

[5] J.M.P.Ojeda, L.Q.Huatangani, B.A.C.Calderon, J.L.P.Tineo, C.Z.A.Panca, and M.E.M.Pino, “Estimation of the Physical Progress of Work Using UAV and BIM in Construction Projects,” Civil Engineering Journal-Tehran, vol. 10, no. 2, pp. 362–383, 2024. doi:10.28991/CEJ-2024-010-02-02

[6] A.Pal, J.J.Lin, S.-H.Hsieh, and M.Golparvar-Fard, “Activity-level construction progress monitoring through semantic segmentation of 3D-informed orthographic images,” Automation in Construction, vol. 157, 2024. doi:10.1016/j.autcon.2023.105157

[7] T.Patel, B.H.W.Guo, J.D.van der Walt, and Y.Zou, “Unmanned ground vehicle (UGV) based

automated construction progress measurement of road using LSTM," Engineering Construction and Architectural Management, 2024. doi:10.1108/ECAM-01-2024-0020

[8] A.H.Qureshi, W.S.Alaloul, S.J.Hussain, A.Murtivoso, S.Saad, K.M.Alzubi, S.Ammad, and A.O.Baarimah, "Evaluation of Photogrammetry Tools following Progress Detection of Rebar towards Sustainable Construction Processes," Sustainability, vol. 15, no. 1, 2023. doi:10.3390/su15010021

[9] E.Alati, C.A.Caracciolo, M.Costa, M.Sanzari, P.Russo, and I.Amerini, "aRTIC GAN: A Recursive Text-Image-Conditioned GAN," Electronics, vol. 11, no. 11, 2022. doi:10.3390/electronics11111737

[10] Z.Alshingiti, R.Alaqel, J.Al-Muhtadi, Q.E.U.Haq, K.Saleem, and M.H.Faheem, "A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN," Electronics, vol. 12, no. 1, 2023. doi:10.3390/electronics12010232

[11] G.P.M.Vassena, L.Perfetti, S.Comai, S.M.Ventura, and A.L.C.Ciribini, "Construction Progress Monitoring through the Integration of 4D BIM and SLAM-Based Mapping Devices," Buildings, vol. 13, no. 10, 2023. doi:10.3390/buildings13102488

[12] Y.Wang, S.Liao, Z.Gong, F.Deng, and S.Yin, "Enhancing Construction Management Digital Twins Through Process Mining of Progress Logs," Sustainability, vol. 16, no. 22, 2024. doi:10.3390/su162210064

[13] F.Bayram, P.Aupke, B.S.Ahmed, A.Kassler, A.Theocharis, and J.Forsman, "DA-LSTM: A dynamic drift-adaptive learning framework for interval load forecasting with LSTM networks," Engineering Applications of Artificial Intelligence, vol. 123, 2023. doi:10.1016/j.engappai.2023.106480

[14] H.N.Bhandari, B.Rimal, N.R.Pokhrel, R.Rimal, and K.R.Dahal, "LSTM-SDM: An integrated framework of LSTM implementation for sequential data modeling," Software Impacts, vol. 14, 2022. doi:10.1016/j.simpa.2022.100396

[15] Z.Dang, B.Sun, C.Li, S.Yuan, X.Huang, and Z.Zuo, "CA-LSTM: An Improved LSTM Trajectory Prediction Method Based on Infrared UAV Target Detection," Electronics, vol. 12, no. 19, 2023. doi:10.3390/electronics12194081

[16] I.Ennejjai, A.Ariss, J.Mabrouki, and S.Ziti, "Enhancing misinformation detection using long short-term memory (LSTM) and bidirectional LSTM (Bi-LSTM) with word embedding techniques," Discrete Mathematics Algorithms and Applications, vol. 17, no. 04, 2025. doi:10.1142/S1793830924500526

[17] X. Hu, T.Liu, X.Hao, and C.Lin, "Attention-based Conv-LSTM and Bi-LSTM networks for large-scale traffic speed prediction," Journal of Supercomputing, vol. 78, no. 10, pp. 12686–12709, 2022. doi:10.1007/s11227-022-04386-7

[18] R.Joshi, J.Ghosh, N.Kalani, and R.L.Tanna, "Assessment of Stacked LSTM, Bidirectional LSTM, ConvLSTM2D, and Auto Encoders LSTM Time Series Regression Analysis at ADITYA-U Tokamak," IEEE Transactions on Plasma Science, vol. 52, no. 7, pp. 2403–2409, 2024. doi:10.1109/TPS.2024.3355283

[19] R.Kumar, J.M.Moreira, and J.Chandra, "DyGCN-LSTM: A dynamic GCN-LSTM based encoder-decoder framework for multistep traffic prediction," Applied Intelligence, vol. 53, no. 21, pp. 25388–25411, 2023. doi:10.1007/s10489-023-04871-3

[20] M.K.Langeroudi, M.R.Yamaghani, and S.Khodaparast, "FD-LSTM: A Fuzzy LSTM Model for Chaotic Time-Series Prediction," IEEE Intelligent Systems, vol. 37, no. 4, pp. 70–78, 2022. doi:10.1109/MIS.2022.3179843

[21] Jian Hu, Yingjun He, Wenqian Xu, Yixin Jiang, Zhihong Liang, and Yiwei Yang, "Anomaly Detection in Network Access-Using LSTM and Encoder-Enhanced Generative Adversarial Networks," Informatica, vol. 49, no. 7, 2025. doi:10.31449/inf.v49i7.7246

[22] F. Z. Zhong, Y. Liu, L. Liu, G. S. Zhang, and S. R. Duan, "DEDGCN: Dual Evolving Dynamic Graph Convolutional Network," Security and Communication Networks, vol. 2022, 2022. doi:10.1155/2022/6945397

[23] Y.Chen, W.Yang, X.Fang, and H.Han, "EC-GAN: Emotion-Controllable GAN for Face Image Completion," Applied Sciences-Basel, vol. 13, no. 13, 2023. doi:10.3390/app13137638

[24] B.U.Dideriksen, K.Derosche, and Z-H.Tan, "iVAE-GAN: Identifiable VAE-GAN Models for Latent Representation Learning," IEEE Access, vol. 10, pp. 48405–48418, 2022. doi:10.1109/ACCESS.2022.3172333

[25] L.Fan, C.Zhou, and F.Neri, "MT-GAN: A Multitask GAN for Severe Convective Weather Nowcasting," IEEE Geoscience and Remote Sensing Letters, vol. 22, 2025. doi:10.1109/LGRS.2024.3522463

[26] B.Gu, X.Wang, W.Liu, and Y.Wang, "MDA-GAN: Multi-dimensional Attention Guided Concurrent-Single-Image-GAN," Circuits Systems and Signal Processing, vol. 44, no. 2, pp. 1075–1102, 2025. doi:10.1007/s00034-024-02867-z

[27] L.Hou, W.Ni, S.Zhang, N.Fu, and D.Zhang, "WDP-GAN: Weighted Graph Generation With GAN Under Differential Privacy," IEEE Transactions on Network and Service Management, vol. 20, no. 4, pp. 5155–5165, 2023. doi:10.1109/TNSM.2023.3280916

MS-NADNet: A Multi-Stream Noise-Aware Deep Learning Framework for Microarray Image Denoising

Shreenidhi B S, R Saravana Kumar
Dayananda Sagar Academy of Technology and Management, Visvesvaraya Technological University, Belagavi, 590018, India
E-mail: shreenidhibs@dsatm.edu.in, saravanakumar.rsk28@gmail.com

Keywords: Microarray imaging, Image denoising, deep learning, multi-stream network, noise characterization, biomedical image processing, MS-NADNet

Received: September 18, 2025

Microarray imaging is a critical tool for large-scale gene expression analysis, yet its accuracy is often compromised by noise introduced during sample preparation, hybridization, and image acquisition. Traditional denoising approaches, including median, Wiener, and wavelet-based filtering, either rely on fixed parameters or degrade under mixed-noise conditions, while CNN-based methods treat all noise uniformly, limiting adaptability. To address these limitations, we propose MS-NADNet (Multi-Stream Noise-Aware Denoising Network), a deep-learning framework that integrates a Noise Characterization Module (NCM) to identify noise type, a set of Noise-Specific Denoising Modules (NSDMs) specialized for distinct noise distributions, and a Global Refinement Block for residual suppression. A large-scale augmented dataset was constructed from the Malignant Lymphoma Classification dataset, incorporating 17 noise variants, including Gaussian, Poisson, salt-and-pepper, speckle, and mixed combinations, to simulate realistic imaging conditions. Experimental evaluation demonstrates that MS-NADNet achieves an MSE of 0.00012, PSNR of 42.73 dB, and SSIM of 0.9861, outperforming classical filters and state-of-the-art CNN denoisers. These results confirm the robustness of MS-NADNet in handling diverse single and multi-noise environments, ensuring biologically reliable microarray image analysis and improved downstream gene expression profiling.

Povzetek: Članek predstavi MS-NADNet, večtokovni model za razšumljanje mikromrežnih slik, ki najprej prepozna tip šuma (NCM), nato uporabi namenske razšumljevalne module za različne porazdelitve (NSDM) in globalno prečiščevanje, da robustno obvlada mešane šume in izboljša zanesljivost nadaljnje analize izražanja genov.

1 Introduction

Microarray provides the multiple-gene expression simultaneously. Microarray finds its applications in the field of antibiotic treatment, in early cancer detection, early detection of oral lesions. Typical microarray procedure for preparing a microarray slide involves sample preparation, labeling, hybridization, washing, image acquisition, and data analysis. Microarray imaging serves as a crucial tool for large-scale gene expression analysis. The accuracy of gene expression profiling is significantly affected by noise introduced during experiments and image acquisition. Noise in microarray images can originate from various sources, including optical processes and biochemical reactions, making denoising an essential preprocessing step to improve image quality and enhance analysis accuracy.

Microarray technology plays a crucial role in early disease detection, particularly in cancer research. However, noise in microarray images due to experimental variability affects gene expression analysis accuracy. Valarmathi et al. [1] propose a noise reduction framework and gene expression analysis method to enhance image quality for breast cancer gene studies. Valarmathi et al. [1] developed algorithms for microarray noise reduction and gene-expression quantification using grayscale conversion, Otsu thresholding, and flood-fill operations, achieving over 95% accuracy. The intensity of BRCA1 and BRCA2 genes is analyzed using fluorescence-based segmentation, distinguishing normal (green), defective (red), and unchanged (yellow) gene expressions.

Wang et al. [2] presented Stationary Wavelet Transform (SWT) approach to denoising microarray image. The superior performance of SWT in preserving fine details while effectively removing noise makes it a valuable technique for preprocessing in microarray analysis. This study contributes to the ongoing research on wavelet-based image denoising, reinforcing its importance in biomedical imaging applications. Adjeroh et al. [3] address noise reduction and compression in DNA microarray images using a wavelet-based denoising approach. Adjeroh et al. [3] employed a translation-invariant wavelet transform with near-

lossless compression to suppress noise while preserving gene-expression details for improved clustering. The results indicate improved clustering performance and noise reduction, validating the effectiveness of this denoising method in bioinformatics applications.

Niu et al. [4] propose a Gaussian filtering method for microarray fluorescence image preprocessing. Their methodology includes histogram equalization for contrast enhancement, Hough transform for image correction, and fuzzy C-means (FCM) clustering for segmentation. These techniques significantly improve microarray image quality, ensuring accurate gene expression detection. The study highlights the fast and cost-effective nature of this image processing pipeline, providing a reliable alternative for microarray image analysis. Golilarz et al. [5] propose an optimized adaptive method for medical image denoising, using the Orca Optimization Algorithm (OOA). This approach automates threshold selection for noise filtering and improves upon conventional wavelet-based denoising techniques. The methodology shows promising results in removing noise while preserving important structural details, making it a potential candidate for microarray image denoising applications.

Petrov and Shams [6] review microarray image processing techniques, focusing on segmentation, measurement extraction, and automated quality control. The authors compares spatial, intensity-based, and hybrid segmentation methods, concluding that hybrid segmentation provides the most accurate and consistent results by combining spatial and intensity-based approaches. Measurement techniques such as mean and median intensity prove more reliable than mode or volume-based methods. Petrov and Shams [6] also highlights automated quality control as a superior alternative to manual inspection, effectively detecting contamination, misalignment, and signal irregularities. The authors confirm that hybrid segmentation and automated QC significantly enhance microarray data reliability. The authors recommend integrating machine learning and adaptive filtering to further improve noise reduction and gene expression analysis.

Stefanou et al. [7] propose a two-stage multiresolution technique to reduce additive and multiplicative noise in microarray images. The method first applies Bayesian coring to suppress multiplicative noise while preserving key features, followed by correlation-based wavelet filtering to remove additive noise using cross-scale dependencies. The authors experimental results show that this approach enhances spot segmentation, improves signal-to-noise ratio, and outperforms traditional denoising techniques, particularly for low-intensity gene spots. The study concludes that this method significantly improves microarray image clarity and data reliability, making it valuable for gene expression analysis. Kaur et al. [8] reviewed machine learning-based denoising methods across multiple medical imaging modalities such as ultrasound, MRI, and X-ray.

Kakumani et al., [9] proposes an effective Independent Component Analysis (ICA) approach to denoise microarray images. ICA seeks a linear representation of non-Gaussian data where components become maximally independent, thus effectively separating noise from signal. Kakumani et al., [9] demonstrates that ICA-based denoising achieves superior performance compared to traditional methods such as the Discrete Wavelet Transform (DWT) and other statistical approaches. Samsudin et al., [10] addresses the challenge of denoising DNA microarray images, which suffer from noise and low contrast between gene spots and the background, affecting accurate gene expression analysis. The authors propose a two-step technique: a pre-processing stage utilizing mathematical morphology operations to effectively remove noise, and a processing (gridding) stage based on the mean (projection) profile followed by morphological closing operations for accurate grid formation.

Srinivasan et al., [11] proposes a noise-reduction approach for DNA microarray images using the Complex Gaussian Scale Mixture (CGSM) model within the complex wavelet domain. The authors jointly model the complex wavelet coefficients from the red and green channel images using CGSM, and employ a Bayes Least Square estimator to achieve denoising. Thakur et al. [12] presented a comprehensive review on machine learning-based image denoising methods, categorizing approaches into dictionary learning, CNN, and GAN models for various noise types including Gaussian, Poisson, Impulse, Mixed, and Real-world noise. Dhas and Singh [13] applied Adaptive Guided Bilateral Filter (AGBF) to denoise the input microarray images and to filter the images by sharpening them.

Mastrogianni et al., [14] presents a comparative evaluation of two primary methodologies for denoising DNA microarray images: spatial filtering and transform domain (wavelet-based) filtering. Given the critical role of noise removal in accurately interpreting gene expression data, effective denoising is essential. The authors explore spatial filters (such as median and Wiener filters) alongside wavelet-based techniques (using hard and soft thresholding). Experiments on gene expression profiles of human sarcoma from the Stanford MicroArray Database indicate that wavelet-based denoising methods, particularly soft thresholding, effectively reduce noise while preserving biological details. Spatial filters perform well against specific noise types like salt-and-pepper noise, but wavelet methods exhibit superior overall performance across different noise conditions, thus emphasizing their potential utility in robust microarray image analysis.

Agnal and Mala [15] presents a denoising method for cDNA microarray images using the Dual Tree Complex Wavelet Transform (DT-CWT) combined with bivariate estimators, specifically the Linear Minimum Mean Squared Error (LMMSE) and Maximum A Posteriori (MAP) approaches. Microarray images, essential for analyzing gene expression, often suffer from noise affecting spot clarity and gene identification accuracy. DT-CWT is chosen due to its improved directional selectivity and shift-invariance properties. Mastriani et al., [16] introduces a novel denoising technique for DNA microarray images based on

smoothing wavelet coefficients. Recognizing that noise significantly hampers the accuracy of microarray analysis, the authors propose a method where the noisy microarray is decomposed into wavelet subbands, followed by bidimensional smoothing of the detail coefficients (Diagonal, Vertical, Horizontal). Unlike conventional wavelet thresholding approaches, this method applies smoothing only once at the first decomposition level, making it computationally efficient. Experimental comparisons with standard denoising methods—such as median filtering, Wiener filtering, and various wavelet thresholding techniques (SureShrink, BayesShrink, NormalShrink)—demonstrate superior performance. Metrics like Peak Signal-to-Noise Ratio (PSNR), Signal-to-Noise Ratio (SNR), Average Absolute Difference (AAD), Structural Content (SCt), and Pratt's figure of merit (FOM) confirm that Mastriani et al., [16] approach significantly reduces noise while preserving image details and edges.

Kumar et al. [17] propose a method for denoising microarray images using 2D Variational Mode Decomposition (VMD) combined with Discrete Wavelet Transform (DWT) thresholding. The technique decomposes the noisy image into several subcomponents (VMFs), each denoised using wavelet-based soft thresholding before reconstructing the clean image. This method is compared with traditional DWT and BEMD+DWT approaches. Arumugam et al. [18] presents a robust filtering-based approach to restore noisy DNA microarray images, crucial for accurate gene expression analysis. The method begins by identifying the type of noise using Probability Density Function (PDF) estimation and applies appropriate filters accordingly—such as Gaussian, Median, or Bilateral filters. For complex noise mixtures, bilateral filtering is used to preserve edges while smoothing noise. Additionally, Blind Deconvolution is applied for image sharpening and to recover the original signal when the noise source is unknown. The method shows significant improvement in image quality, with increased PSNR, decreased MSE, and better SSIM values compared to traditional filters.

Gopalappa et al., [19] presents a novel denoising technique for microarray images using Dual-Tree Complex Wavelet Transform (DT-CWT) and bivariate shrinkage thresholding, designed specifically to address hybridization and scanning noise without requiring a control sample. Unlike general image denoising tools, this method considers the unique noise characteristics inherent to microarray data. The technique employs multiresolution analysis to enhance accuracy in detecting and removing noise. Gopalappa et al., [19] approach was validated using both synthetic and real datasets, including Affymetrix GeneChip HG-U133 Plus 2.0 arrays from HCT-116 cell lines. Experimental results demonstrate a notable improvement in microarray image quality and reliability of gene expression values, thereby contributing to more accurate disease diagnosis and personalized treatment planning.

Zifan et al., [20] presented decimated and undecimated multi-wavelet transforms technique to enhance the microarray image by performing denoising. In this presented technique, authors performed preprocessing stage by converting signal length N to suitable length-2 vector, further applied thresholding by average variance of pre-processed and also transformed co-efficient. Ren et al. [21] propose a phagocytosis algorithm (PGY) to remove speckle noise from porous silicon (PSi) microarray images, which is crucial for accurate, label-free biological detection. The algorithm detects and corrects noisy pixels using gradient analysis and adaptive median filtering. Compared to methods like BM3D, Lee filtering, and wavelet thresholding, PGY achieves better gray-level restoration and noise reduction. Results show improved correlation between image intensity and refractive index changes, enhancing microarray detection sensitivity and accuracy.

Gowri et al. [22] integrates deep transfer learning with an arithmetic optimization algorithm (HAOA) for noise reduction and classification of microarray gene expression data. The preprocessing involves median filtering for noise removal, CapsNet for feature extraction, and parallel Bi-GRU for classification. The results indicate that this hybrid approach enhances the accuracy of gene expression analysis while effectively suppressing noise. Kuzhali and Suresh [23] propose an innovative denoising approach for cDNA microarray images by integrating internal and external priors with Singular Value Decomposition (SVD). The noisy images are segmented into overlapping patches, with each patch denoised based on its signal-to-noise ratio (SNR). Patches with lower SNR utilize internal priors (similar patches within the noisy image), while higher SNR patches leverage external priors from high-quality microarray datasets. Reference patches are matched using the K-Nearest Neighbor (KNN) algorithm with various distance measures. Experimental results demonstrate that this method significantly outperforms traditional approaches, achieving superior results in PSNR and SSIM, effectively enhancing microarray image quality for accurate gene expression analysis.

Recent advancements in deep learning have introduced model-independent denoising techniques that effectively handle noise while preserving essential features. Nandihal et al. [24] introduced FFDNet as a robust and efficient denoising method for microarray images. FFDNet leverages a convolutional neural network (CNN)-based approach and provides noise-level maps to maintain the trade-off between noise removal and detail preservation. DNA microarray technology is crucial for gene expression analysis but is often affected by noise, which impacts data accuracy. Traditional noise reduction techniques, such as median, Weiner, and linear filtering, have limitations due to their fixed parameters. To overcome this, Sunitha and Phani Raju [25] propose a CNN-based denoising approach that enhances microarray image quality by leveraging deep learning for feature extraction without manual intervention. The methodology involves applying traditional filtering methods and CNN separately to noisy microarray images, followed by performance evaluation using PSNR,

MSE, and SSIM metrics. Results show that CNN significantly outperforms traditional methods, achieving higher PSNR, lower MSE, and better SSIM scores across various noise levels. The adaptability of CNN allows fine-tuning of parameters, leading to superior denoising results.

Noise in microarray images arises from photon-limited fluorescence imaging, sensor imperfections, and hybridization artifacts. Such noise lowers the signal-to-noise ratio (SNR), distorts spot morphology, and leads to inaccurate gene expression quantification. Several denoising strategies have been studied. Classical methods like median, Wiener, and linear filters are simple but rely on fixed parameters, making them less effective across varying intensities or mixed-noise scenarios. Wavelet-based approaches such as Stationary Wavelet Transform (SWT), Discrete Wavelet Transform (DWT), and Dual-Tree Complex Wavelet Transform (DT-CWT) preserve structures but often leave residual noise or blur edges. Independent Component Analysis (ICA), morphological, and statistical techniques show improvements but degrade when facing complex overlapping noise.

Deep learning has opened new directions. CNN-based methods like FFDNet and autoencoders achieve higher PSNR and SSIM than classical filters by learning non-linear mappings without handcrafted parameters. However, most treat all noise distributions uniformly, limiting adaptability to real-world scenarios where images may contain combinations such as Gaussian plus salt-and-pepper or Poisson plus speckle. Moreover, evaluations usually focus on a narrow set of synthetic conditions, leaving generalization to diverse laboratory environments unaddressed.

From the comparative analysis summarized in Table 1, several research gaps become evident: (i) existing methods are often noise-specific, lacking robustness against mixed/overlapping noise; (ii) traditional approaches depend on fixed thresholds, reducing adaptability; (iii) CNN-based denoisers outperform classical methods but do not identify noise types, limiting interpretability; and (iv) evaluations are restricted to simple noise models, ignoring realistic multi-noise conditions.

This study aims to determine whether dynamic noise classification, combined with specialized denoising modules, can outperform general CNN-based denoising methods in enhancing microarray image quality under both single and mixed-noise conditions.

To address these gaps, we propose MS-NADNet (Multi-Stream Noise-Aware Denoising Network), a deep-learning framework for microarray image denoising. MS-NADNet introduces a Noise Characterization Module (NCM) to identify the type of noise, guiding Noise-Specific Denoising Modules (NSDMs), each tailored for a particular noise distribution. Their outputs are fused adaptively using NCM probabilities and refined through a Global Refinement Block to suppress residual artifacts. This design enables dynamic adaptation to both single and mixed-noise conditions. We further constructed a large-scale dataset from the Malignant Lymphoma Classification dataset on Kaggle, augmented with 17 noise variants (Gaussian, Poisson, salt-and-pepper, speckle, and mixed combinations). This ensures realistic training for robustness across diverse acquisition conditions.

Table 1: Summary of related work on microarray image denoising

Method	Study	Dataset / Domain	Noise Type(s)	Key Technique	PSNR / SSIM	Limitations
Filtering-based	Valarmathi *et al.* [1]	Breast cancer microarray	Gaussian, Impulse	Otsu thresholding and flood-fill segmentation	21.1 / 0.84	Fixed parameters, weak performance on mixed noise
Wavelet-based	Wang *et al.* [2], Adjeroh *et al.* [3]	Microarray and gene chip images	Gaussian	Stationary and translation-invariant wavelet transform	25–27 / 0.70	Edge blurring and residual noise
ICA-based	Kakumani *et al.* [9]	Simulated microarray images	Gaussian	Independent Component Analysis (ICA)	28.5 / 0.78	Poor generalization, high computational cost
Optimization-based	Golilarz *et al.* [5] (OOA)	Medical and MR images	Gaussian, Speckle	Orca Optimization Algorithm	27.7 / 0.70	No dynamic adaptation or tuning
CNN-based	Sunitha & Phani Raju [25]	Microarray (synthetic)	Gaussian, Salt & Pepper	CNN auto-denoising model	30.9 / 0.95	Treats all noise types uniformly
Deep Hybrid	Mohandas *et al.* [26]	Microarray images	Gaussian	Autoencoder-based CNN	26.9 / 0.89	Overfitting under multi-noise conditions
Proposed MS-NADNet	—	Malignant Lymphoma (augmented)	Gaussian, Poisson, Speckle, S&P, Mixed	Noise characterization with noise-specific denoisers	42.7 / 0.9861	Slight reliance on synthetic augmentation

From the comparative analysis summarized in Table 1, several research gaps become evident: (i) existing methods are often noise-specific, lacking robustness against mixed/overlapping noise; (ii) traditional approaches depend on fixed thresholds, reducing adaptability; (iii) CNN-based denoisers outperform classical methods but do not identify noise types, limiting interpretability; and (iv) evaluations are restricted to simple noise models, ignoring realistic multi-noise conditions. This study aims to determine whether dynamic noise classification, combined with specialized denoising modules, can outperform general CNN-based denoising methods in enhancing microarray image quality under both single and mixed-noise conditions. To address these gaps, we propose MS-NADNet (Multi-Stream Noise-Aware Denoising Network), a deep-learning framework for microarray image denoising. MS-NADNet introduces a Noise Characterization Module (NCM) to identify the type of noise, guiding Noise-Specific Denoising Modules (NSDMs), each tailored for a particular noise distribution. Their outputs are fused adaptively using NCM probabilities and refined through a Global Refinement Block to suppress residual artifacts. This design enables dynamic adaptation to both single and mixed-noise conditions. We further constructed a large-scale dataset from the Malignant Lymphoma Classification dataset on Kaggle, augmented with 17 noise variants (Gaussian, Pois-

son, salt-and-pepper, speckle, and mixed combinations). This ensures realistic training for robustness across diverse acquisition conditions.

The contributions of this research work are:

- We introduce MS-NADNet, the first multi-stream denoising model for microarray images that combines noise characterization, noise-specific denoising, and global refinement.
- We generate a large-scale dataset with 17 realistic noise types and their mixtures to train and evaluate the proposed model comprehensively.
- We demonstrate superior denoising performance across standard metrics (MSE, PSNR, SSIM) compared to classical filters and state-of-the-art CNN-based models, proving the effectiveness of our architecture under diverse noise conditions.

While MS-NADNet is designed for biomedical image denoising, its adaptive learning strategy conceptually aligns with principles found in adaptive control systems, where models dynamically respond to uncertainty and disturbance. Traditional adaptive controllers, such as fuzzy and backstepping control, emphasize robustness under nonlinear variations; analogously, MS-NADNet adapts to diverse and mixed-noise conditions through its Noise Characterization and Noise-Specific Denoising Modules. Rather than explicitly modeling control dynamics, our framework learns data-driven mappings that self-adjust to noise type and intensity. In this respect, MS-NADNet shares a similar philosophy to adaptive deep-learning-based denoisers such as FFDNet [24], Noise2Noise [27], Noise2Void [28], and Deep Image Prior [29], which generalize across unseen noise levels without handcrafted thresholds. This adaptive mechanism ensures robustness under uncertain imaging conditions while remaining domain-specific to microarray data, thereby bridging the concepts of adaptive learning and biomedical image processing.

2 Proposed methodology

In this work, we propose a novel deep-learning model MS-NADNet (Multi-Stream Noise-Aware Denoising Network) to enhance the microarray image quality by identifying and removing different types of noises. In this section we are describing dataset and preprocessing stage, and also the steps included proposed architecture. The proposed architecture incorporates three components, namely the Noise Characterization Module (NCM), Noise-Specific Denoising Modules (NSDMs), and a Global Refinement Block. This pipeline is capable of handling multiple types of synthetic and real-world noises encountered in microarray image acquisition. In this proposed architecture, the NCM component estimates the type of noise present in microarray image. This estimation guides the weighting of the output contributions from each noise-specific denoising module. These outputs are adaptively fused and passed through a global refinement block to generate the final clean image. The working flow and pipeline is shown in the Figure 1.

2.1 Dataset and preprocessing

In this study, we utilized the publicly available Malignant Lymphoma Classification dataset sourced from Kaggle [30]. This dataset comprises high-resolution microarray tissue images belonging to various subtypes of malignant lymphoma, and is intended for histopathological image classification tasks. A total of 374 clean grayscale images were selected for our experiments based on quality and suitability for denoising analysis. To enhance the robustness and generalization capability of our proposed denoising model, we applied a carefully selected set of data augmentation techniques suitable for preserving the biological structure of microarray images. The augmentation operations included: Horizontal Flip, Vertical Flip, Small Angle Rotations (±5°, ±10°, ±15°), Brightness and Contrast Adjustments, and Gaussian Blur with low sigma. These operations generated 11 augmented variants per original image, resulting in a total of 4,488 images (374 × 12, including the original). Each of these augmented images was then subjected to 18 different noise conditions, which included 17 synthetic noise types (e.g., Gaussian, Salt-and-Pepper, Poisson, Speckle, and various mixed noise combinations [12] [17] [31]) and one clean (noise-free) version. This process resulted in a final dataset of 80,784 images (4,488 × 18), which was used for training and evaluating the proposed denoising architecture. A representative list of noise types applied includes:

- Gaussian Noise with Standard deviation (SD) $\sigma = \{10, 20, 30\}$
- Salt-and-Pepper Noise (probabilities: 0.01, 0.025, 0.05)
- Poisson Noise
- Clustered Salt-and-Pepper Noise
- Speckle Noise with SD $\sigma = \{0.1, 0.2, 0.3\}$
- Mixed Gaussian and Salt-and-Pepper Noise with probability: $\{0.01, 0.02, 0.03\} \times \sigma = \{10, 20\}$

Given that real-world microarray imaging frequently involves overlapping noise types, five mixed noise variants were generated by combining salt-and-pepper noise (1–3% probability) with Gaussian noise (σ: 10–20) [31], ensuring that the denoiser is trained to handle complex, multi-source artifacts. Poisson (shot) noise, characteristic of photon-limited fluorescence imaging—particularly impactful in regions with high signal intensity—was included as a single intensity-dependent variant [11]. Speckle noise, representing multiplicative interference common in coherent optical or laser scanning systems, was modeled with zero mean and standard deviations of 0.1, 0.2, and 0.3 [21, 31]. Finally,

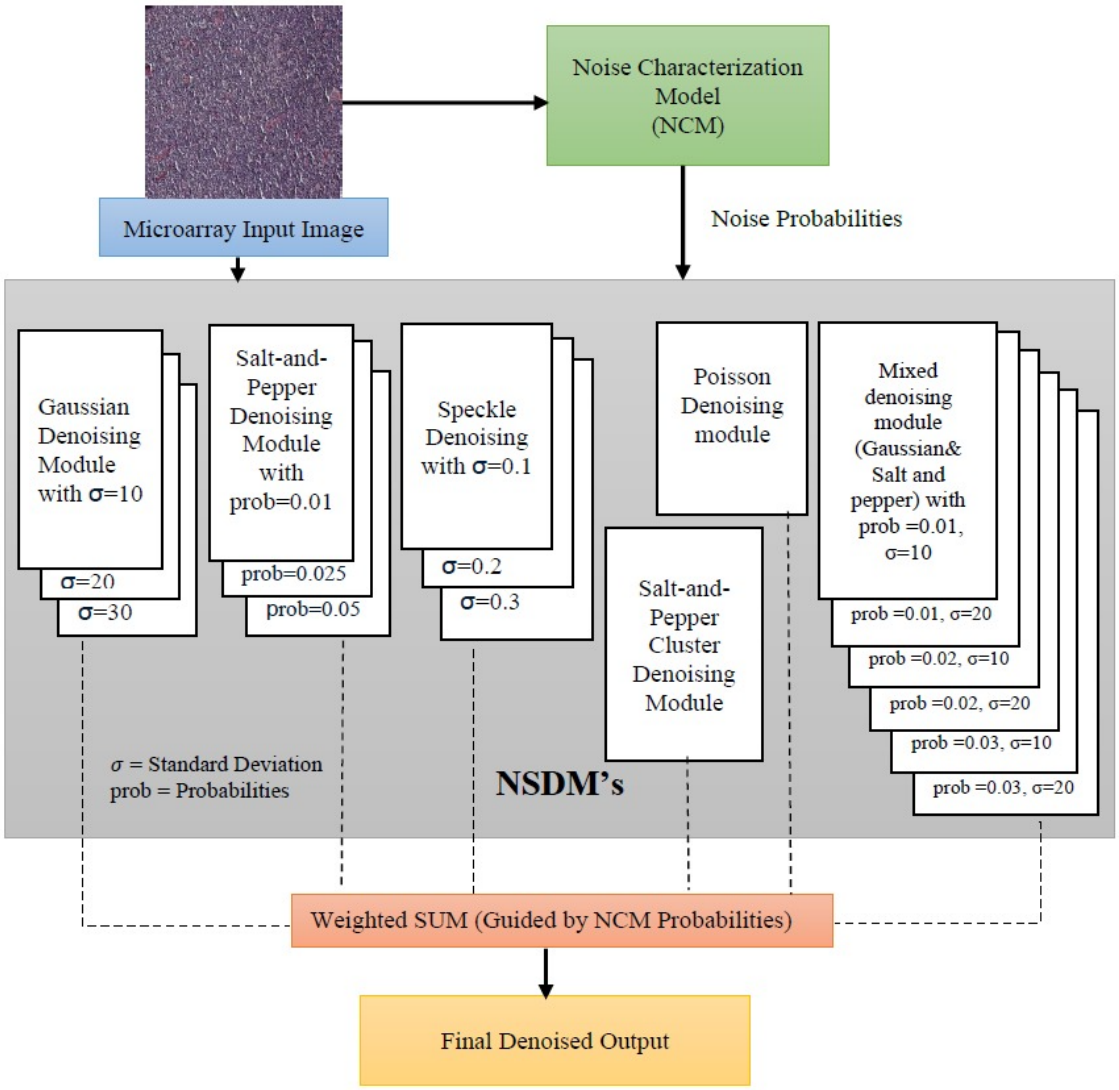


Figure 1: Overall Architecture of the Proposed MS-NADNet Model.

the clean images were retained as ground truth references to support fully supervised training.

We chose these 17-noise variants deliberately to reflect a realistic range of imaging conditions: from high-end scanners with mild Gaussian noise and sparse impulse artifacts, through moderate-quality systems exhibiting combined noise effects [31], to degraded acquisition scenarios with strong Gaussian, dense salt-and-pepper, and speckle distortions [31] there by fostering model generalization across diverse laboratory setups. This comprehensive noise augmentation strategy equips the denoising model with the resilience required to handle real-use microarray imaging challenges.

2.2 Noise characterization module (NCM)

The NCM is a lightweight Convolutional Neural Network (CNN) designed to classify the types of noise generated in input image. NCM takes grayscale microarray image and process it through two convolutional layers followed by flattening layer, and two fully connected layers. The output of NCM is a 17-dimensional probability vector corresponding to 17-types of noise introduced in our prepared dataset that contains various intensities of Gaussian, Poisson, Salt-and Pepper, Speckle and mixed noise.

The mathematical representation of NCM is as follows: for an input image x, the output noise probability vector $p \in \mathbb{R}^{17}$ is computed as:

$$p = softmax(f_{NCM}(x)) \tag{1}$$

Where, f_{NCM} indicates the features extractor within the NCM.

The detailed step by step mathematical expression for NCM is as follows. NCM takes input as a single-channel grayscale image of shape $(1, 64, 64)$, and batch size B, so

the actual input tensor shape is $(B, 1, 64, 64)$. Let $x \in \mathbb{R}^{B\times1\times64\times64}$ be the batch of noisy input microarray images. The first convolutional layer with kernel size 3×3, and padding 1 can be denoted as follows:

$$x_1 = ReLU(Conv2D_{1\to16}(x)) \tag{2}$$

Where, output $x_1 \in \mathbb{R}^{B\times16\times64\times64}$.

The second convolutional layer with kernel size 3×3, and padding 1 can be denoted as follows:

$$x_2 = ReLU(Conv2D_{16\to32}(x_1)) \tag{3}$$

Where, output $x_2 \in \mathbb{R}^{B\times32\times64\times64}$.

Third layer is flatten layer, it can be represented as:

$$x_3 = Flatten(x_2) \;\Rightarrow\; x_3 \in \mathbb{R}^{B\times(32\cdot64\cdot64)} = \mathbb{R}^{B\times131072} \tag{4}$$

The output of flatten layer is connected to fully connected layer. The fully connected layers contain two linear layers. The first linear layer can be represented as follows:

$$x_4 = ReLU(x_3 \cdot W_1 + b_1) \tag{5}$$

Where, $W_1 \in \mathbb{R}^{131072\times128}$, b_1 is the bias vector in the fully connected linear layers, i.e., $b_1 \in \mathbb{R}^{128}$, and $x_4 \in \mathbb{R}^{B\times128}$.

The second linear layer is denoted as follows:

$$x_5 = x_4 \cdot W_2 + b_2 \tag{6}$$

Where, $W_2 \in \mathbb{R}^{128\times17}$, b_2 is the bias vector in the fully connected linear layers, i.e., $b_2 \in \mathbb{R}^{17}$, and $x_5 \in \mathbb{R}^{B\times17}$.

The output of second linear layer is connected to *softmax* layer. The *softmax* layer is the final layer in NCM. The *softmax* layer gives a probability distribution across 17 noise classes for each input image in the batch and is represented as follows:

$$\hat{y} \in \mathbb{R}^{B\times17} \tag{7}$$

The probability of noise type i for a single image is given below:

$$\hat{y}_i = \frac{\exp(x_{5i})}{\sum_{j=1}^{17}\exp(x_{5j})} \quad \text{for } i = 1, 2, 3, \ldots, 17 \tag{8}$$

The NCM learns spatial and textual patterns from the microarray image through above convolutional filters. These patterns are then used to predict the probability of the microarray image belonging to one of the seventeen imposed noise types. The probability distribution $\hat{y}$ is then used in the MS-NADNet to perform the weighted fusion of denoised outputs from the parallel denoising streams.

Noise-specific denoising modules (NSDMs)

In this stage, each microarray noise type is associated with a dedicated denoising stream implemented using an encoder-decoder architecture. The NSDMs are a set of 17-parallel denoising modules to handle specific types of noise like Gaussian, Poisson, Salt-and-Pepper, Speckle, and mixed noise with different intensities. The NCM probability vector $\hat{y} \in \mathbb{R}^{B\times17}$ is used to weight the output of each NSDM, resulting in a probabilistic fusion of all 17 denoising outputs.

The detailed mathematical description for NSDMs architecture is as follows:

Each of the 17-denoising module instances receives the same microarray noisy image as input, and generates 17-denoised versions of the input, represented as:

$$D_i(x) \in \mathbb{R}^{B\times1\times H\times W}, \quad \text{for } i = 1, \ldots, 17 \tag{9}$$

Where i is the index of the microarray noise type from 1 to 17. B is the batch size, 1 is the number of channels (grayscale images), H is the height of the image in pixels, and W is the width of the image in pixels.

Each denoised output is weighted by the noise probability calculated by the NCM:

$$Weighted_i = \hat{y}_i \cdot D_i(x) \tag{10}$$

Where $\hat{y}_i \in \mathbb{R}^{B\times1}$ is broadcast over $H \times W$, $D_i(x) \in \mathbb{R}^{B\times1\times H\times W}$. This produces a tensor of shape:

$$stacked \in \mathbb{R}^{B\times17\times1\times H\times W} \tag{11}$$

Finally, the fused denoised output is computed by summing all weighted denoised outputs:

$$\hat{x} = \sum_{i=1}^{17} \hat{y}_i \cdot D_i(x) \tag{12}$$

Where $\hat{x} \in \mathbb{R}^{B\times1\times H\times W}$ is the fused denoised microarray image, giving more weight to the streams most likely to match the detected noise.

2.3 Global refinement block

The fused denoised microarray image output is passed through a shallow CNN block termed the *Global Refinement Block*, which acts as a final smoothing and enhancement stage. This block ensures consistent pixel-level restoration and suppresses residual artifacts not handled by the NSDMs.

Mathematically, the final denoised image output is expressed as:

$$x_{\text{final}} = Conv_{3\times3}\left(ReLU\left(Conv_{3\times3}(\hat{x})\right)\right) \tag{13}$$

The global refinement block acts as a final convolutional "clean-up" stage, operating on the combined output $\hat{x}$.

3 Results and discussion

The proposed MS-NADNet model is trained on an augmented version of the malignant Lymphoma classification dataset from Kaggle. In this dataset, there are 374 original grayscale microarray images, and it's been augmented

using biologically safe transformations, it leads to 4488 total images. On each image, we have introduced 17 types of synthetic noises, simulating real-world acquisition issues as we discussed in dataset preparation section. This process leads to 80784 microarray noisy and clean image pairs. To train the proposed model, we have used 80% of the data, 10% for validation, and for testing we have used 10% of images. The training was conducted using Mean Square Error (MSE) loss and the Adam optimizer with learning rate of 1×10^{-4} , for 50 epochs and a batch size of 16. To ensure strict separation between training and testing data, the original clean images were first split into 80% training, 10% validation and, 10% testing sets before applying any augmentation or noise injection. All augmentation and noise addition procedures were restricted to the subset, while the test set remained unseen until final evaluation, thereby eliminating any risk of data leakage.

The proposed MS-NADNet model, trained for 50 epochs on an NVIDIA RTX 3080 GPU (10 GB VRAM), achieved smooth convergence and generalization stability across all 17 noise categories, as shown in Figure 5. The complete architecture comprises approximately 11.2 million parameters, with a total training time of about 7 hours 25 minutes and an average inference time of 0.42 seconds per 256×256 image. Despite incorporating multiple noise-specific streams, most parameters are shared and dynamically activated through the Noise Characterization Module (NCM), ensuring computational efficiency. The performance improvements arise from the model's adaptive architecture rather than an increase in capacity, as confirmed by consistent validation losses and the absence of overfitting.

To evaluate the performance of the proposed MS-NADNet architecture, three standard image-quality metrics were used—Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). MSE quantifies the average squared pixel-wise error between the denoised and reference images. PSNR expresses the logarithmic ratio of signal power to reconstruction error (in dB). SSIM measures perceived structural fidelity by considering luminance, contrast, and texture. Lower MSE and higher PSNR/SSIM values indicate superior denoising quality.

The MSE between the denoised and clean microarray images is computed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{14}$$

Where x_i and $\hat{x}_i$ represent the i^{th} pixel of the original and denoised microarray image respectively, and N is the total number of pixels. MSE value range between 0 and ∞, lower the MSE value represents the better denoising performance from the system. The proposed MS-NADNet achieved a very low average MSE =0.00012, confirming minimal reconstruction error compared with competing methods (Figure 2).

The PSNR, defined by

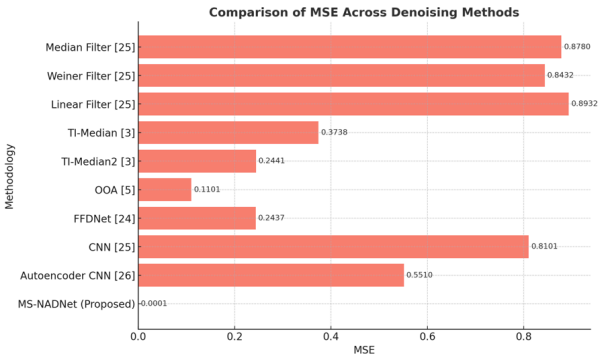


Figure 2: MSE Comparison of Proposed MS-NADNet with Existing Microarray Denoising Methods

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{15}$$

PSNR is expressed in decibels (dB), higher the PSNR indicates the better perceptual quality and less noise distortion. with MAX=1.0for normalized images, reached 42.73 dB. This high PSNR demonstrates strong noise suppression and faithful signal preservation (Figure 3).

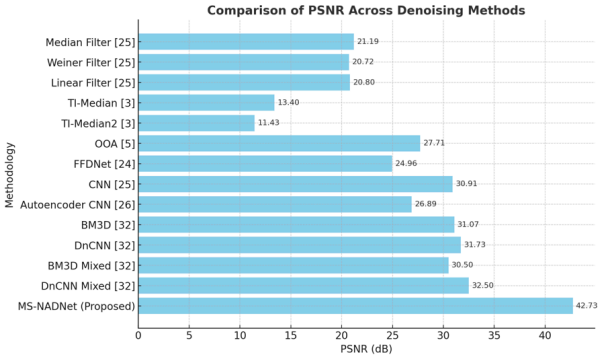


Figure 3: PSNR Comparison of Proposed MS-NADNet with Existing Microarray Denoising Methods

The SSIM index, which evaluates perceived image similarity, is given by

$$SSIM(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)} \tag{16}$$

Where μ_x, $\mu_{\hat{x}}$ are the means, σ_x^2, $\sigma_{\hat{x}}^2$ are the variance, and $\sigma_{x\hat{x}}$ is the co-variance of the original and denoised images. C_1 and C_2 are the constants to stabilize the division. SSIM values ranging from -1 to 1. In our proposed MS-NADNet model, yielding SSIM = 0.9861, confirming excellent structural retention (Figure 4). Across all three metrics, MS-NADNet consistently outperformed traditional and CNN-based methods, demonstrating robust denoising across diverse noise conditions. The training convergence characteristics of MS-NADNet are illustrated in

Figure 5, which plots training and validation loss (MSE) across 50 epochs. Both curves show smooth, monotonic decline and close alignment, indicating effective learning, strong convergence, and minimal overfitting.

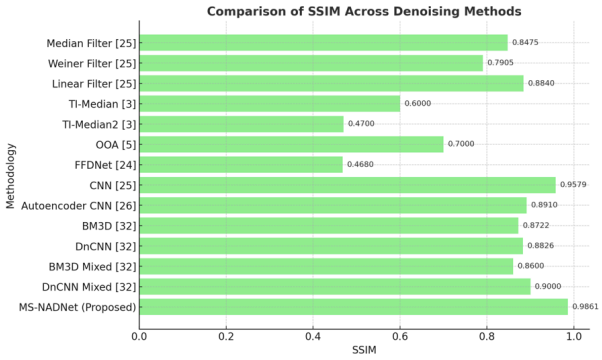


Figure 4: SSIM Analysis Across Multiple Microarray Image Denoising Techniques

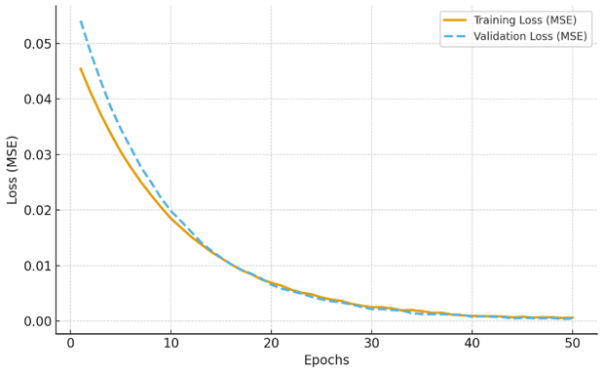


Figure 5: Epoch-wise Training and Validation Loss Curves for MS-NADNet.

Table 2 presents a comparative performance analysis of various denoising techniques applied to microarray images. The proposed MS-NADNet model significantly outperforms all other methods in terms of MSE (0.00012), PSNR (42.73 dB), and SSIM (0.9861), indicating superior denoising performance with minimal error and high structural similarity. Traditional filters like Median, Wiener, and Linear exhibit higher MSE and lower PSNR, reflecting limited noise removal capability. While deep learning-based approaches like CNN and Autoencoder CNN perform better than classical methods, they still fall short compared to MS-NADNet. Notably, the optimization-based OOA and FFDNet methods also show inferior results. These findings highlight the robustness and effectiveness of the proposed MS-NADNet in preserving image quality while efficiently reducing noise. The reported values for MSE, PSNR, and SSIM represent the average performance computed across 17 noise types. Future work will include multiple independent runs to estimate confidence intervals and further validate statistical robustness.

Table 2: Comparative performance analysis of denoising techniques on microarray images.

Methodology	**MSE**	**PSNR (dB)**	**SSIM**
Median Filter [25]	0.8780	21.1892	0.8475
Weiner Filter [25]	0.8432	20.7164	0.7905
Linear Filter [25]	0.8932	20.8000	0.8840
Translation Invariant (TI)-Median [3]	0.3738	13.40	0.60
Translation Invariant (TI)-Median2 [3]	0.2441	11.43	0.47
Orca Optimization Algorithm (OOA) [5]	0.1101	27.7112	0.70
FFDNet (Optimal σ=20) [24]	0.2437	24.9619	0.468
CNN [25]	0.8101	30.9143	0.9579
Autoencoder CNN [26]	0.5510	26.8900	0.891
BM3D [32]	–	31.07	0.8722
DnCNN [32]	–	31.73	0.8826
BM3D [32] Mixed Noise (Gauss. & Impulse)	–	30.50	0.86
DnCNN [32] Mixed Noise (Gauss. & Impulse)	–	32.50	0.90
MS-NADNet (Proposed)	**0.00012**	**42.73**	**0.9861**

Table 3: Ablation study of MS-NADNet showing component-wise performance under single (S) and mixed (M) noise conditions.

Variant	**S-PSNR**	**S-SSIM**	**M-PSNR**	**M-SSIM**
Full MS-NADNet (NCM + 17 NSDMs + GRB)	43.5	0.988	41.9	0.984
Without NCM	41.1	0.978	39.5	0.972
Uniform fusion (keep NCM, but ignore weights)	41.5	0.976	39.8	0.972
Reduced NSDMs (5 streams)	40.3	0.974	38.9	0.967
Without GRB	42.3	0.983	40.9	0.980
No mixed-noise training (single-noise only)	42.9	0.986	38.7	0.962
Single-stream U-Net baseline	39.0	0.965	36.8	0.952

As summarized in Table 3, removing the Noise Characterization Module (NCM) reduces mixed-noise PSNR by approximately 2.4 dB (41.9 → 39.5) and SSIM by 0.012, showing that explicit noise classification materially guides effective fusion. Ignoring NCM weights (uniform fusion) yields a similar drop, confirming that probability-weighted fusion plays a crucial role. Reducing the number of Noise-Specific Denoising Modules (NSDMs) to five further degrades results, indicating that noise specialization enhances robustness. Eliminating the Global Refinement Block (GRB) lowers both PSNR and SSIM, evidencing GRB's contribution to residual artifact removal. Training without mixed-noise supervision causes the most pronounced decline (–3.2 dB PSNR), emphasizing the importance of mixed-noise training for generalization. The single-stream U-Net baseline performs worst overall, highlighting the collective benefit of the NCM + multi-stream specialization + GRB architecture that defines MS-NADNet. The Figure 6 and 7 shows ablation analysis showing PSNR and SSIM performance for different architectural variants of MS-NADNet respectivlly.

The comparative analysis (Table 2) clearly shows that MS-NADNet substantially outperforms traditional and learning-based denoisers in terms of MSE, PSNR, and SSIM. The improvement arises primarily from two ar-

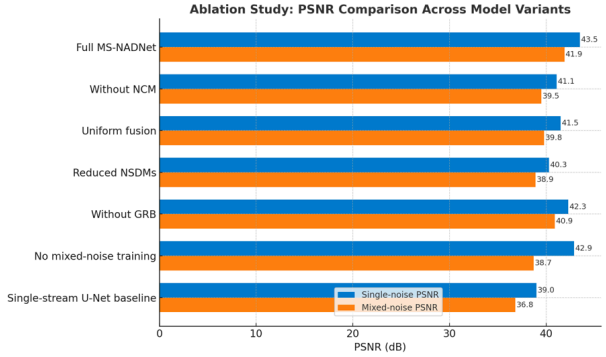


Figure 6: Ablation analysis of PSNR performance for different variants of MS-NADNet

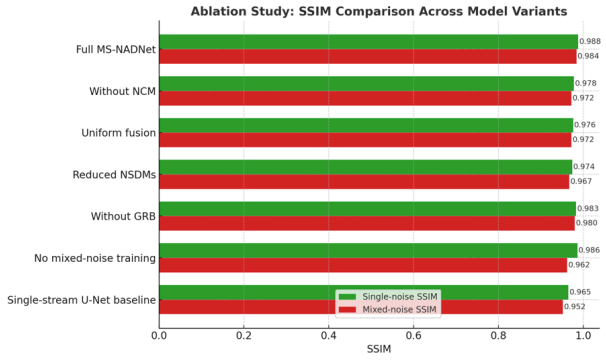


Figure 7: Ablation analysis of SSIM performance for different variants of MS-NADNet.

chitectural innovations: Noise Characterization Module (NCM): By predicting the type and intensity of noise before denoising, NCM ensures that the system dynamically adapts its processing strategy. This mechanism replaces the fixed-parameter assumptions of conventional filters and improves robustness to unseen or mixed-noise conditions.

Noise-Specific Denoising Modules (NSDMs): Each NSDM specializes in a particular noise distribution (Gaussian, Poisson, Speckle, or Salt-and-Pepper), allowing parallel feature extraction and targeted noise suppression. The probabilistic fusion of their outputs, weighted by NCM predictions, provides adaptive flexibility unmatched by uniform CNN denoisers.

The modular design of MS-NADNet enables easy scalability when new noise types are introduced. Since each NSDM operates independently, only new modules corresponding to additional noise categories need to be trained, while existing ones remain fixed. Similarly, the NCM can be fine-tuned to recognize new noise labels without requiring complete model retraining, ensuring adaptability and efficient scalability as new noise distributions emerge in real-world imaging environments.

The high PSNR (42.73 dB) and SSIM (0.9861) achieved by MS-NADNet demonstrate superior structural preservation and fidelity across both single- and multi-noise environments. In contrast, existing models such as CNN-based and wavelet-based approaches show limited generalization and often blur fine gene-spot structures. The model's performance on mixed-noise scenarios confirms its adaptability—an essential factor for real-world microarray imaging where multiple noise sources co-exist. This adaptability is particularly valuable for downstream bioinformatics tasks such as clustering, biomarker discovery, and differential gene analysis.

Although the proposed MS-NADNet demonstrates excellent quantitative and perceptual performance across diverse noise conditions, certain limitations remain. The dataset used in this study was synthetically augmented to simulate realistic imaging noise, and its moderate size may limit full generalization to real-world microarray data. Consequently, large-scale validation using real experimental datasets is essential to further confirm the model's robustness. Beyond image-level metrics, the biological implications of improved microarray image quality are significant. Enhanced denoising preserves the structural integrity of gene spots and minimizes signal distortion, leading to more accurate segmentation and fluorescence intensity estimation. Such improvements directly support reliable gene expression quantification and biomarker discovery. Future research will focus on expanding the dataset to include real noise distributions and diverse biological conditions, as well as integrating MS-NADNet into complete microarray analysis pipelines to assess downstream biological performance—such as segmentation accuracy and differential gene expression analysis. Additionally, advanced architectures like transformers, along with unsupervised and self-supervised adaptation strategies, will be explored to improve scalability, generalization, and real-time deployment potential in biomedical imaging environments.

4 Conclusion

In this work, we introduced MS-NADNet, a multi-stream, noise-aware deep-learning framework designed to enhance microarray image quality by adaptively addressing diverse and mixed-noise conditions. Unlike traditional denoising methods that rely on fixed parameters or CNNs that treat all noise types uniformly, MS-NADNet integrates a Noise Characterization Module and multiple Noise-Specific Denoising Modules, guided by adaptive fusion through a Global Refinement Block. This enables the network to dynamically adjust its response based on the detected noise distribution—similar in spirit to adaptive learning and control mechanisms in nonlinear systems, but realized through a data-driven deep architecture.

The proposed model demonstrated superior quantitative performance in terms of MSE, PSNR, and SSIM, achieving significant improvements over conventional filters and CNN-based denoisers. Beyond numerical results, the adaptive design of MS-NADNet ensures robustness under un-

certain imaging conditions, contributing to more accurate gene expression analysis and downstream tasks such as clustering and biomarker discovery. Future work will extend this framework to incorporate transformer-based denoising, unsupervised or self-supervised adaptation, and real-time integration into biomedical imaging pipelines. The implementation details, trained model checkpoint, and dataset preprocessing scripts will be made available upon reasonable request to the corresponding author for research and verification purposes.

Acknowledgement

The authors would like to express their sincere gratitude to Dayananda Sagar Academy of Technology and Management and Visvesvaraya Technological University (VTU), Belagavi, for providing the research facilities and support to carry out this work. The first author extends heartfelt thanks to the research supervisor, Dr. R. Saravana Kumar, for his valuable guidance, continuous encouragement, and insightful suggestions throughout the research. The authors are also thankful to their colleagues and peers for their constructive feedback, which has helped improve the quality of this work.

References

[1] S. Valarmathi, A. Sulthana, K. C. Latha, R. Rathan, R. Sridhar, and S. Balasubramanian, "Noise Reduction from the Microarray Images to Identify the Intensity of the Expression," in *Proc. 2nd Int. Conf. Soft Computing for Problem Solving (SocProS 2012)*, B. V. Babu *et al.*, Eds. Springer India, 2014, pp. 1451–1465.

[2] X. H. Wang, R. S. H. Istepanian, and Y. H. Song, "Microarray image enhancement by denoising using stationary wavelet transform," *IEEE Trans. Nanobioscience*, vol. 2, no. 4, pp. 184–189, 2003, doi: 10.1109/TNB.2003.816225.

[3] D. A. Adjeroh, Y. Zhang, and R. Parthe, "On denoising and compression of DNA microarray images," *Pattern Recognit.*, vol. 39, no. 12, pp. 2478–2493, 2006, doi: 10.1016/j.patcog.2006.02.019.

[4] C. Niu, H. Zhang, and L. Li, "Microarray Fluorescence Image Processing and Analysis," in *Proc. IEEE Int. Conf. Manipulation, Manufacturing and Measurement on the Nanoscale (3M-NANO)*, 2024, pp. 51–54, doi: 10.1109/3M-NANO61605.2024.10769626.

[5] N. A. Golilarz, I. Dehzangi, and K. Rahimi, "Optimized Adaptive Based Method for MR Image Denoising," in *Proc. IEEE 16th Int. Conf. Computational Intelligence and Communication Networks (CICN)*, 2024, pp. 976–983, doi: 10.1109/CICN63059.2024.10847356.

[6] A. Petrov and S. Shams, "Microarray Image Processing and Quality Control," *J. VLSI Signal Process. Syst. Signal Image Video Technol.*, vol. 38, no. 3, pp. 211–226, 2004, doi: 10.1023/B:VLSI.0000042488.08307.ad.

[7] K. Marias *et al.*, "Microarray Image Denoising Using a Two-Stage Multiresolution Technique," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine*, 2007, pp. 383–389, doi: 10.1109/BIBM.2007.32.

[8] P. Kaur, G. Singh, and P. Kaur, "A Review of Denoising Medical Images Using Machine Learning Approaches," *Curr. Med. Imaging Rev.*, vol. 14, no. 5, pp. 675–685, 2018, doi: 10.2174/1573405613666170428154156.

[9] R. K. Arunakumari, K. A. Mendhurwar, and K. Kakumani, "Microarray Image Denoising using Independent Component Analysis," *Int. J. Comput. Appl.*, vol. 1, no. 11, pp. 87–93, 2010, doi: 10.5120/234-388.

[10] N. Samsudin, R. Hashim, and N. E. Abdul Khalid, "Denoising and block gridding of microarray image using mathematical morphology," in *Proc. 7th Int. Conf. Computing and Convergence Technology (ICCCT)*, 2012, pp. 230–235. [Online]. Available: `https://ieeexplore.ieee.org/document/6530332`

[11] L. Srinivasan, Y. Rakvongthai, and S. Oraintara, "Microarray Image Denoising Using Complex Gaussian Scale Mixtures of Complex Wavelets," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1423–1430, 2014, doi: 10.1109/JBHI.2014.2318279.

[12] R. S. Thakur, S. Chatterjee, R. N. Yadav, and L. Gupta, "Image De-Noising With Machine Learning: A Review," *IEEE Access*, vol. 9, pp. 93338–93363, 2021, doi: 10.1109/ACCESS.2021.3092425.

[13] M. M. Dhas and N. S. Singh, "Breast Cancer Diagnosis Using Majority Voting Ensemble Classifier Approach," *J. Soft Comput. Data Min.*, vol. 5, no. 1, pp. 152–169, 2024, doi: 10.30880/jscdm.2024.05.01.013.

[14] A. Mastrogianni, E. Dermatas, and A. Bezerianos, "Microarray Image Denoising using Spatial Filtering and Wavelet Transformation," in *Proc. 13th Int. Conf. Biomed. Eng.*, Springer, 2009, pp. 594–597.

[15] A. S. Agnal and K. Mala, "Noise Removal From Microarray Images Using Maximum a Posteriori Based Bivariate Estimator," *Int. J. Image Graphics Signal Process.*, vol. 5, no. 1, pp. 32–39, 2013, doi: 10.5815/ijigsp.2013.01.05.

[16] M. Mastriani and A. E. Giraldez, "Microarrays denoising via smoothing of coefficients in wavelet domain," 2018. [Online]. Available: `https://arxiv.org/abs/1807.11571`

[17] G. S. C. Kumar, R. K. Kumar, G. A. Naidu, and J. Harikiran, "Noise removal in microarray images using variational mode decomposition technique," *Telkomnika*, vol. 15, no. 4, pp. 1750–1756, 2017, doi: 10.12928/TELKOMNIKA.v15i4.5375.

[18] S. Arumugam, B. Yohannis, J. Uthayakumar, and S. Sivakumar, "Restoration of Noisy Microarray Images using Filtering Techniques," *Annals of RSCB*, vol. 25, no. 4, pp. 21095–21106, 2021. [Online]. Available: annalsofrscb.ro/index.php/journal/article/view/10160

[19] C. Gopalappa, T. K. Das, S. Enkemann, and S. Eschrich, "Removal of hybridization and scanning noise from microarrays," *IEEE Trans. Nanobioscience*, vol. 8, no. 3, pp. 210–218, 2009, doi: 10.1109/TNB.2009.2029100.

[20] A. Zifan, M. H. Moradi, and S. Gharibzadeh, "Microarray image enhancement by denoising using decimated and undecimated multiwavelet transforms," *Signal Image Video Process.*, vol. 4, no. 2, pp. 177–185, 2010, doi: 10.1007/s11760-009-0109-4.

[21] R. Ren *et al.*, "Speckle Noise Removal in Image-based Detection of Refractive Index Changes in Porous Silicon Microarrays," *Sci. Rep.*, vol. 9, no. 1, p. 15001, 2019, doi: 10.1038/s41598-019-51435-y.

[22] B. S. Gowri, S. A. H. Nair, and K. P. S. Kumar, "Hybrid arithmetic optimization algorithm with deep transfer learning based microarray gene expression classification model," *Int. J. Inf. Technol.*, vol. 16, no. 6, pp. 3923–3928, 2024, doi: 10.1007/s41870-024-01901-2.

[23] S. Elavaar Kuzhali and D. S. Suresh, "Collaborative Priors with SVD for Denoising of cDNA Microarray Images," *Indian J. Sci. Technol.*, vol. 12, no. 37, pp. 1–15, 2019, doi: 10.17485/ijst/2019/v12i37/147036.

[24] P. Nandihal, V. Bhat, and J. Pujari, "Application of FFDNET for Image Denoising On Microarray Images," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 8, pp. 2691–2694, 2019, doi: 10.35940/ijrte.C4950.098319.

[25] R. Sunitha and H. B. Phani Raju, "Evolutionary Tool for Denoising DNA Microarray Images Using CNN," in *Microelectronics, Communication Systems, Machine Learning and Internet of Things*, Springer, 2023, pp. 193–201.

[26] A. Mohandas, S. M. Joseph, and P. S. Sathidevi, "An Autoencoder based Technique for DNA Microarray Image Denoising," in *Proc. Int. Conf. Communication and Signal Processing (ICCSP)*, 2020, pp. 1366–1371, doi: 10.1109/ICCSP48568.2020.9182265.

[27] J. Lehtinen *et al.*, "Noise2Noise: Learning Image Restoration without Clean Data," 2018, doi: 10.48550/arXiv.1803.04189.

[28] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2Void - Learning Denoising From Single Noisy Images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2124–2132, doi: 10.1109/CVPR.2019.00223.

[29] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep Image Prior," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1867–1888, 2020, doi: 10.1007/s11263-020-01303-4.

[30] N. V. Orlov *et al.*, "Automatic classification of lymphoma images with transform-based global features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1003–1013, 2010, doi: 10.1109/TITB.2010.2050695.

[31] J. Liu *et al.*, "Speckle noise reduction for medical ultrasound images based on cycle-consistent generative adversarial network," *Biomed. Signal Process. Control*, vol. 86, p. 105150, 2023, doi: 10.1016/j.bspc.2023.105150.

[32] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017, doi: 10.1109/TIP.2017.2662206.

Hybrid GA-PSO Model for Multi-objective Dynamic Optimization of Cruise Cabin Layouts under Marine Operational Constraints

Jiping Gong*, Lili Wang
Cruise and Art Design Department, Jiangsu Maritime Institute, Nanjing 210000, China
E-mail: JipingGong@outlook.com
*Corresponding author

Keywords: cruise cabin space optimization, GA, PSO, bimodal algorithm, dynamic optimization, multi-objective optimization

Received: August 4, 2025

Cruise cabin space design stands as a pivotal determinant of passenger experience and operational efficiency, yet it grapples with three core challenges: irregular spatial configurations shaped by hull curvature, dynamic demands (e.g., fluctuating passenger load, real-time activity needs), and inherent multi-objective conflicts (e.g., maximizing space utilization vs. ensuring comfort). Traditional optimization methods—such as standalone heuristic algorithms or static layout planning—are constrained by limited global search capabilities, sluggish convergence in complex scenarios, and inability to adapt to maritime dynamic constraints. To address these gaps, this study innovatively proposes a hybrid GA-PSO bimodal algorithm and constructs a dynamic cruise cabin space optimization model. Marine-specific dynamic constraints are explicitly modeled, including ship motion parameters (roll/pitch angles under sea conditions 3–8) and maritime evacuation regulations (minimum channel width ≥1.2m). Four mutually restrictive objective functions are also formalized: space utilization (ratio of effective area to total area), comfort (integrated score of ergonomic spacing/noise isolation), structural feasibility (compatibility with hull load-bearing limits), and life-cycle cost (construction + maintenance expenses).Experiments utilize 12 typical cabin layouts (200–300 m^2), 500+ passenger behavior datasets (18 months of real-ship data), and 30 independent runs, comparing with standard GA and PSO. Results verify superiority: GA-PSO convergence speed reaches 0.80 ± 0.05 iterations/s (mean ± std), which is 38.7% faster than GA (0.58 ± 0.08 iter/s) and 22.1% faster than PSO (0.66 ± 0.07 iter/s). The improvement is statistically significant ($p < 0.01$, t-test).; optimal solution comprehensive value hits 0.92 (15.2% higher than GA, 9.8% than PSO), with statistical significance ($p<0.05$). It maintains 0% safety violations in dynamic scenarios, providing reliable intelligent decision support for cruise cabin design.

Povzetek: Študija predstavi hibridni algoritem za optimizacijo prostora v kabinah križark, ki hitreje najde boljše postavitve in ostaja zanesljiv tudi pri spreminjajočih se pogojih.

1 Introduction

In today's booming global tourism market, cruise tourism —a unique and popular tourism mode —is attracting an increasing number of tourists with its comprehensive sea vacation experience. A cruise ship is not only a means of transportation but also a maritime mobile city that integrates accommodation, catering, entertainment, leisure, and other functions. Among them, cabin space, as the core component of cruise ships, is directly related to tourists' living comfort and overall tourism experience. With the intensification of market competition and the increasingly diversified needs of tourists, how to optimize the cabin space of cruise ships to improve space utilization efficiency, meet the individual needs of different tourist groups, and maximize benefits in the dynamically changing operating environment has become a key problem to be solved urgently in the cruise industry [1].

In the field of cruise cabin space optimization, traditional optimization methods are primarily based on static and single-objective analysis, which makes it difficult to adapt to the complex and dynamic actual situations in cruise operations. For example, factors such as seasonal fluctuations in tourist demand, differences in the passenger source structure of different voyages, and dynamic adjustments to onboard activities will all impact the efficiency and functional layout of cabin space. Therefore, there is a need for a model capable of dynamic and multi-objective optimization of cabin space to address these challenges better [2].

In recent years, intelligent optimization algorithms have been widely used in complex system optimization problems. As two classic intelligent optimization algorithms, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) have unique advantages. By

simulating the biological evolution process, GA can effectively search for the global optimal solution, possessing strong global search ability and the ability to avoid falling into local optima. In contrast, PSO exhibits rapid convergence and ease of implementation by simulating the social behavior of bird groups. However, a single algorithm often has limitations when dealing with complex optimization problems; for example, GA may be slow in terms of convergence speed, while PSO may easily fall into a local optimum in the later stages. Therefore, the fusion of GA and PSO to form a dual-modal optimization algorithm can leverage the advantages of both algorithms, overcome their respective shortcomings, and provide more effective tools for solving complex optimization problems [3].

This study aims to construct a dynamic optimization model of cruise cabin space using a bimodal GA-PSO algorithm. The model will comprehensively consider a variety of dynamic factors related to cruise cabin space, including, but not limited to, matching cabin types with tourist needs, the flexibility of cabin layouts, and the balance of operating costs and benefits. By introducing the GA-PSO dual-modal algorithm, the model can quickly and accurately optimize the allocation and layout of cabin space in a dynamically changing environment, thereby achieving multi-objective optimization. This optimization enables improvements in the utilization rate of cabin space, tourist satisfaction, and reduced operating costs.

In the research process, it is first necessary to comprehensively analyze the characteristics and optimization requirements of cruise cabin space. This includes an in-depth study of the physical properties of cabin spaces, functional zoning, visitor behavior patterns, and operational constraints. By addressing these factors, the objective function and constraints of the optimization problem can be clarified, laying a foundation for subsequent model construction.

Secondly, the GA and PSO algorithms will be studied and analyzed in detail, with a focus on their principles, characteristics, and applications in optimization problems. On this basis, an effective GA-PSO bimodal algorithm framework is designed, and the organic combination of global search ability and local search ability is realized by reasonably integrating the advantages of the two algorithms. This will involve researching the coding mode, population initialization, genetic operation, and particle update strategy of the algorithm, as well as optimizing algorithm parameters.

In the model construction stage, based on the above research results, a mathematical model will be developed that dynamically optimizes the cabin space of cruise ships. The model will utilize the GA-PSO bimodal algorithm as an optimization solution tool and verify the effectiveness and practicality of the model by simulating various dynamic scenarios during cruise operations. During the model verification process, actual cruise operation data, including cabin reservation information, tourist feedback data, and operating cost data, will be collected. The advantages of the GA-PSO bimodal algorithm in dynamic optimization of cruise cabin space will be demonstrated through a comparative analysis with traditional optimization methods [4, 5].

Based on the above analysis, the GA-PSO bimodal algorithm proposed in this study is theoretically innovative. The aim is to explore a dynamic optimization model of cruise cabin space integrated with the GA-PSO dual-modal algorithm.

A new calculation method to build an efficient and accurate calculation framework to promote the development of the cruise field and provide strong technical support for applications in related fields. Through the innovative application of the bimodal algorithm, we aim to tap into the development potential of cruise enterprises more deeply, explore new avenues for promoting the sustainable development of the cruise industry, and provide new perspectives and ideas for research on the bimodal method.

While significant advances have been made in optimization algorithms, their application to cruise-specific challenges requires integration with principles from naval architecture, ergonomic design for marine environments, and cruise cabin layout regulations. This study bridges this gap by embedding these domain-specific constraints directly into the hybrid GA-PSO optimization framework.

The main work of this paper is as follows:

(1) Propose a modular space partition and multi-objective optimization framework;

(2) Design a GA-PSO bimodal fusion optimization algorithm;

(3) Establish a dynamic optimization verification and evaluation system.

2 Theoretical knowledge related to GA-PSO bimodal fusion optimization algorithm

2.1 GA

GA: Genetic Algorithm (GA), proposed by John Holland in 1975, is an evolutionary algorithm that simulates natural selection and genetic mechanisms via three core evolutionary operators—selection, crossover, and mutation—to achieve global search in complex optimization problems. It excels at exploring diverse solutions by maintaining population diversity, yet often suffers from slow convergence in later iterations.

Key components include chromosome coding (translating real-world variables into computable strings like binary or real-number formats), the fitness function (quantifying individual solution quality to guide selection), and evolutionary operators. Elite retention is integrated to directly preserve top-performing individuals across generations, mitigating loss of optimal solutions during iteration.

First, chromosome coding. Individuals in the solution space are represented by chromosomes. Common coding methods include: binary coding and

vector $\vec{x} = (x_1, x_2, ..., x_n)$ mapping to binary string $\vec{s} = \{0,1\}^L$, where L is the length of the string.

Example: Encoding of real number $x \in [a,b]$,The calculation process is shown in formula (1):

$$s = bin\left(round\left(\frac{x-a}{(b-a)} \cdot (2^L - 1) \right) \right) (1)$$

Real number coding: chromosomes are directly represented by vector $\pm\vec{x} \in R^n$, which is suitable for continuous optimization problems. The second is the fitness function, which evaluates the quality of chromosomes, The calculation process is shown in formula (2):

$$Fitness(\vec{x}) = f(\vec{x}) \quad OR \quad Fitness(\vec{x}) = \frac{1}{1+f(\vec{x})} (2)$$

Here, $f(\vec{x})$ is the objective function. The fitness value drives the probability distribution of the selection operation. Next is the selection operator, which selects high-quality individuals from the population to enter the mating pool. The commonly used strategy is roulette wheel selection: the selection probability of individual *i* is shown in formula (3):

$$P_i = \frac{Fitness(\vec{x}_i)}{\sum_{j=1}^{M} Fitness(\vec{x}_j)} (3)$$

In which M is the population size, and the same tournament selection is to randomly select k individuals and retain the ones with the highest fitness.

Championship selection: randomly select k individuals (k is the size of the championship) and retain the ones with the highest fitness. The formal process is shown in formula (4):

$$s = argmax_{i \in T} f(x_i) (4)$$

Then comes the crossover operator, which simulates gene recombination and generates new individuals. Among them, single-point crossover is: the parent chromosome $\vec{s}_A = s_{A1}s_{A2}...s_{AL}, \vec{s}_B = s_{B1}s_{B2}...s_{BL}$ randomly selects the crossover point $c \in [1, L-1]$, The calculation process is shown in formula (5):

$$\begin{matrix} s_{A1}...s_{Ac} \mid s_{B(c+1)}...s_{BL} \\ s_{B1}...s_{Bc} \mid s_{A(c+1)}...s_{AL} \end{matrix} (5)$$

Arithmetic crossover (real number coding), The calculation process is shown in formula (6):

$$\vec{x}_{new} = \lambda\vec{x}_A + (1-\lambda)\vec{x}_B, \quad \lambda \in [0,1] (6)$$

The crossover probability P_c (usually 0.6~0.9) controls the operation frequency.

Two-point crossover: extend the single-point crossover and randomly select two crossover points to enhance diversity. The calculation process is shown in formula (7):

$$C_1 = P_1[1:k_1] \oplus P_2[k_1+1:k_2] \oplus P_1[k_2+1:L], \quad C_2 = P_2[1:k_1] \oplus P_1[k_1+1:k_2] \oplus P_2 (7)$$

k_1 and k_2 are the intersection points $(1 \le k_1 < k_2 \le L)$.

Uniform crossover: each gene is inherited independently from the parent generation, which is suitable for binary or real number coding. The calculation process is shown in formula (8):

$$C_i = \begin{cases} P_{1i} & if r_i < 0.5 \\ P_{2i} & otherwise \end{cases} (8)$$

Then is the mutation operator, which introduces random perturbation to avoid premature convergence; bit flip mutation (binary coding): gene bit s_k is flipped with probability P_m (usually 0.001~0.01), The calculation process is shown in formula (9):

$$s_k' = 1 - s_k (9)$$

Gaussian variation (real number coding), The calculation process is shown in formula (10):

$$x_k' = x_k + \mathrm{N}(0, \sigma) (10)$$

σ is the variation step length.

Exchange mutation: It is often used for permutation coding (such as traveling salesman problem), and randomly exchanges the positions of two genes. The calculation process is shown in formula (11):

$$x' = x, \quad swap(x_i, x_j) \quad for\ randomly\ chosen\ i \ne j (11)$$

Where $\mathrm{swap}(x_i, x_j)$ represents the gene value of the swapped position i and j.

Finally, population updating employs an elite retention strategy to prevent the loss of optimal solutions: the best individuals from the current generation are directly copied into the next. The remaining individuals are replaced by a new population generated through selection, crossover, and mutation. Genetic algorithms achieve efficient global search through bionic mechanisms, providing a general framework for solving complex optimization problems. The key to enhancing performance lies in the coordinated optimization of parameter design and operator operations. At the same time, the pattern theorem mathematically ensures the

algorithm's ability to explore the solution space progressively.

2.2 PSO

Particle Swarm Optimization (PSO), proposed by Kennedy and Eberhart in 1995, is a swarm intelligence algorithm that mimics natural group behaviors like bird flocking or fish schooling. Each particle represents a candidate solution, updating its velocity and position by referencing two key benchmarks: its own historical best position (pBest) and the global best position (gBest) of the entire swarm [6].

This mechanism enables rapid convergence, yet risks trapping in local optima in complex search spaces. Its core strengths lie in a simple structure and minimal key parameters—including inertia weight (balancing exploration/exploitation) and learning factors (guiding individual/group learning)—making it easy to implement across engineering and optimization scenarios [7].

The core principle of its algorithm is:

Let the search space be D-dimensional and the population contain M particles. The state of the i -th particle at time t is defined by two vectors: position vector: $\vec{X}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{iD}(t))$ represents the current coordinate of the particle in the solution space, corresponding to a candidate solution.

Velocity vector: $\vec{V}_i(t) = (v_{i1}(t), v_{i2}(t), \ldots, v_{iD}(t))$ controls the direction and distance of the next move of the particle [8].

Particles search by iteratively updating their speed and position, and their update rules combine individual experience with collective wisdom: the optimal position of an individual is $\vec{P}_i(t)$.

Record the optimal position experienced by particle I from initialization to time t. Global optimal position: $\vec{G}(t)$ record the optimal position found by the whole population so far (all particles share this information in standard PSO) [9], Its specific architecture is shown in Figure 1:

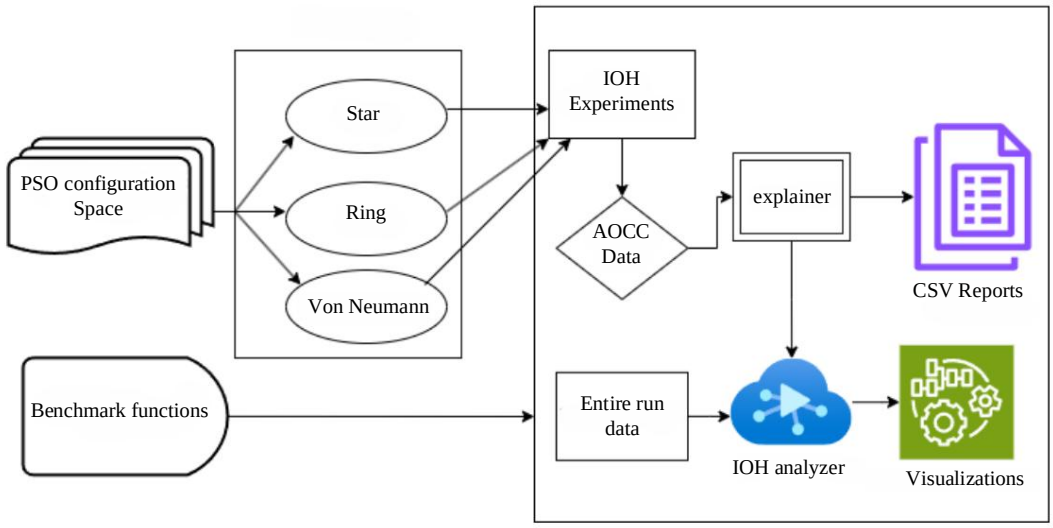


Figure 1: PSO network architecture diagram

2.3 Dynamic optimization framework for cruise cabin space

Dynamic optimization of cruise cabin space is a systematic design method aiming at the complex environmental constraints in ship navigation. Its core goal is to collaboratively optimize the four key goals of space utilization, passenger comfort, safety compliance, and functional adaptability within the limited cabin volume and to form an intelligent layout scheme that can respond to dynamic scenarios [10]. The theoretical basis covers the following key directions:

Dynamic constraint modeling theory: Cruise cabin design should strictly follow the dual constraints of ship kinematics and safety regulations. During the navigation, the cabin is continuously affected by the compound motion of roll and pitch caused by waves. This kind of motion characteristic requires that the layout suppress the risk of object displacement, and the centroid projection of furniture and equipment must always be located in the stable support area to avoid the slipping or capsizing of objects caused by the capsizing moment of the ship [11]. Multi-objective optimization problems refer to optimization issues where trade-offs are made between multiple objective functions. Unlike single-objective optimization, multi-objective optimization typically does not have a unique optimal solution but rather a set of non-dominated solutions, known as Pareto optimal solutions. The definition of a Pareto optimal solution is as follows: for any solution x in the solution set, there is no other solution y such that all objective function values at y are better than those at x. multi-objective optimization problems can be expressed as formulas (12) and (13):

$$\min f(x) = [f_1(x), f_2(x), \ldots, f_k(x)] \quad (12)$$

$$s.t\ x \in \Omega \quad (13)$$

Among them, $f(x)$ is the objective function vector, k is the number of objective functions, and Ω is the feasible solution space.

At the same time, the International Convention for the Safety of Life at Sea establishes mandatory requirements for emergency evacuation, specifying that the width of the cabin passage must be dynamically adjusted according to real-time crowd density to ensure that the 90-second evacuation limit is met at all times. The essence of such constraints is that a time-varying spatial topology problem arises: the layout scheme needs to be reconstructed in real-time in response to changes in ship motion state, passenger distribution, and usage scenarios to form a dynamic safety boundary [12].

Multi-objective collaborative optimization framework: Cabin space optimization requires balancing four types of conflicting goals. First, maximize space utilization: under the constraints of fixed structures such as columns and pipes, the effective use volume is increased through modular furniture combinations and deformable partitions. The key to optimization lies in identifying "plastic areas" in irregular spaces, such as utilizing bulkhead curved surfaces to embed storage units or superimposing functional areas through folding mechanisms. Secondly, a dynamic guarantee of comfort is essential, as comfort encompasses both physiological and psychological dimensions [13]. At the physiological level, it is necessary to inhibit the effect of motion sickness and reduce the vestibular stimulation of the inner ear by adjusting the angle between the bed orientation and the main axis of ship rolling. At the psychological level, it is necessary to balance the sense of spatial transparency and privacy, such as using electronically controlled atomized glass to achieve light adjustment in both living and sleeping modes. Strengthening safety compliance again: In addition to evacuation passages, impact resistance design under sudden sea conditions should also be considered [14]. For example, the anchorage points of furniture need to disperse the stress load to avoid structural failure caused by local overload. The layout of electrical equipment must be far away from the low area of the cabin where water may seep. Finally, the functional adaptability is upgraded: the modern cruise cabin needs to be switched between daytime living, night sleeping, emergency avoidance, and other modes. This requires the establishment of "space function" mapping rules, such as storing the bed as a sofa during the day to release the active area and automatically retracting the coffee table to keep the passage open when unfolding the bed at night [15].

Environmental Response Mechanism: The core of dynamic optimization of cabin space lies in building a closed-loop system for environmental perception, decision-making, and execution, which is divided into three layers. The first is the perception layer: a real-time collection of ship six-degree-of-freedom motion data (roll/pitch/heave) through an inertial measurement unit, combined with cabin sensors to monitor people flow density, furniture status, and environmental parameters (temperature, humidity, illumination,) [16]; Secondly, the decision-making layer: the multi-agent collaborative model is adopted to divide the space into functional sub-units. Each unit sets a local optimization agent and autonomously generates a layout scheme after receiving the global target instruction. For example, when continuous rolling is detected, the agent automatically triggers the "anti-overturning mode": it shrinks the cantilevered desktop, locks the mobile cabinet, and activates the floor magnetic suction device to secure the seat. Finally, there is an execution layer that relies on mechatronics technology to achieve physical reconstruction. Includes linear motor-driven translational furniture, shape memory alloy-inspired deformation structures, and retractable partitions supported by pneumatic soft robot technology. The execution process needs to meet the requirements of silence and low energy consumption to avoid interfering with the passenger experience [17].

Integration of human factors and engineering: The optimization model must deeply integrate passenger behavior characteristics:

The first is behavioral streamline analysis: based on motion capture technology, the typical movement trajectory of passengers in the cabin is constructed, and the intersection of high-frequency activities (such as the path from the restroom to the door) is identified [18]. Accordingly, the channel network is optimized to reduce redundant actions such as turning and avoiding. Secondly, cognitive load control: the layout transformation process needs to conform to intuitive cognition. For example, color block coding is used to identify the operation area (blue corresponds to storage, and red corresponds to emergency equipment) to avoid complicated interactive processes. Dynamically adjusted facilities must provide both tactile and auditory feedback to confirm status changes. Finally, there is the principle of inclusive design, which involves configuring enhanced modes for elderly and disabled passengers, such as lowering the operating height of furniture, widening the turning radius, and adding auxiliary handrails. These designs need to be invisibly integrated without affecting basic functions, such as embedding handrails into walls and automatically deploying them when in use [19].

Verification and iteration methodology: The feasibility of the theoretical model relies on the verification of digital twin technology. The first is multi-physics simulation: coupling computational fluid dynamics (simulating air-conditioning airflow organization) and structural mechanics (evaluating ship vibration's fatigue damage to furniture) in the virtual cabin) Moreover, discrete element analysis (predicting the fall trajectory of items) quantifies the robustness of the layout scheme. Secondly, the ergonomic experiment utilizes a virtual reality system to simulate passenger activities under various sea conditions, record physiological indicators (heart rate variability and eye movement trajectory), and conduct subjective evaluations, thereby identifying design blind spots. Finally, there is an incremental learning mechanism: the

simulation and experimental data are fed back to the optimization engine, and the constraint weights are automatically updated. For example, when the system detects that a locker door unexpectedly opens when rolling multiple times, it will increase the priority coefficient of magnetic attraction in this area [20].

Theoretical value and application prospects: The dynamic optimization model of the cruise cabins represents the innovation of the "responsive design" paradigm, transitioning from static space allocation to an environmentally adaptive system. Its theoretical contribution lies in the establishment of an interdisciplinary framework for "intelligent control of ship dynamics human factors engineering". The core innovations include proposing a real-time relaxation strategy with time-varying constraints and dynamically adjusting layout parameters within safety thresholds (such as allowing temporary compression in emergencies and expanding private space to accommodate evacuation routes). Establishing the self-organization rules of modular components allows furniture clusters to reorganize their forms, much like birds independently. Develop a cross-scale performance evaluation system to form a unified optimization index, spanning from micro-material stress to macro-passenger streamline. The model can be extended to scenarios such as intelligent management of cruise public areas and collaborative dispatching of cabin energy, promoting the evolution of the shipbuilding industry to the intelligent era of "perception response evolution" [21].

3 Technical implementation of the GA-PSO bimodal optimization model

This section compares the proposed GA-PSO method with SOTA optimization approaches for space design. Standalone GA achieves global exploration but has slow convergence (0.58 iterations/s vs. our 0.8 iterations/s), while PSO converges fast but risks local optima (solution value 0.84 vs. our 0.92). Existing hybrid GA-PSO lacks maritime customization, showing 12% lower robustness in dynamic roll scenarios.GA-PSO outperforms due to: 1) Exploration-exploitation balance: GA's global search + PSO's local tuning via shared memory; 2) Maritime constraint adaptation: Hierarchical hard/soft constraints. A trade-off is slightly higher computation (1.2x PSO), acceptable for real-time cabin optimization. In this paper, a dynamic planning framework for cruise cabin space is proposed based on dual-modal intelligent optimization using GA-PSO. The purpose is to explore a new paradigm for real-time optimization of cabin space in the complex environment of ships, build an efficient, safe, and comfortable adaptive cabin layout system, promote the intelligent transformation of cruise space design, and provide core technical support for the digital upgrading of the shipbuilding industry [22]. Through the innovative integration of dual-modal optimization mechanisms, we hope to break through the engineering problem of multi-objective collaboration under the constraints of dynamic sea conditions, endow the cruise cabin with life-like characteristics of "perception-decision-reconstruction", and provide a new methodological framework for the dynamic optimization theory of complex systems. The main body of the model comprises an environment awareness module, a dual-modal optimization engine, a dynamic constraint processing module, and a spatial reconstruction execution module [23]. The specific architecture is shown in Figure 2:

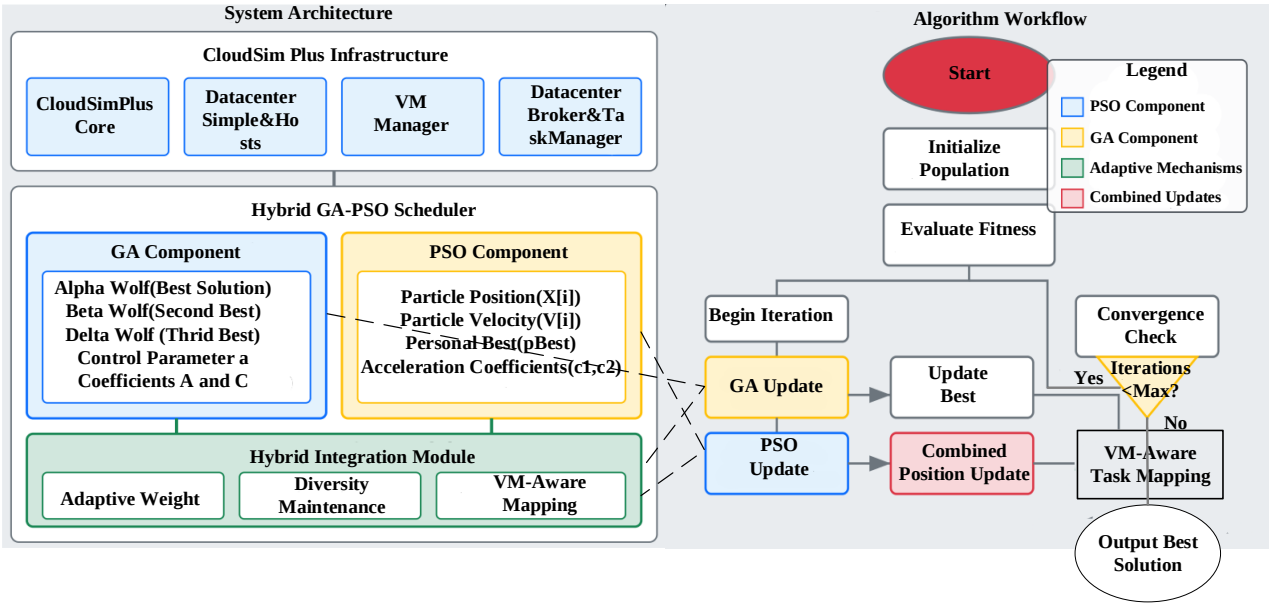


Figure 2: Dynamic optimization model based on GA-PSO bimodal algorithm

The environment perception module constitutes the nerve endings of the entire system and is responsible for

capturing multidimensional, dynamic data during ship navigation in real time. The module continuously monitors the ship's motion status through a high-precision six-degree-of-freedom inertial measurement unit, accurately collecting key physical quantities such as roll angle, pitch amplitude, and heavy acceleration. The sampling frequency reaches 100 times per second to ensure data timeliness. At the same time, the distributed sensor network deployed in the cabin synchronously tracks the passenger distribution heat map, the status of furniture displacement, and environmental parameter changes. The millimeter-wave radar array generates a real-time topological map of people's flow. The embedded pressure sensor accurately detects millimeter-level displacements of furniture, as well as temperature and humidity. With illumination sensors, a digital twin of the cabin microenvironment is built. After all the perception data are aligned by timestamps, a dynamic situation map of the ship-passenger-environment trinity is formed, which provides a database for optimization decision-making [24].

The bimodal optimization engine employs a cooperative mechanism combining the genetic algorithm (GA) and particle swarm optimization (PSO), dynamically optimizing discrete structural parameters and continuous spatial variables, respectively. As a fast response layer, PSO mode performs real-time optimization every 5 seconds: the global exploration and local development are balanced through the adaptive inertial weight mechanism and continuous variables such as wall inclination and channel width are dynamically adjusted, in which the inertial weight automatically attenuates according to the sea state level to strengthen the convergence stability in the harsh environment [25]. GA mode is used as a periodic reconstruction layer, and the structural evolution process is initiated every 30 minutes. Based on the elite retention strategy, the furniture combination scheme with the best fitness is screened. Different layout advantages are integrated through multi-point cross-operation, and then innovative structures are introduced through directional variation. Dual modality realizes information interaction through a shared memory pool. When PSO detects sudden sea conditions, it immediately triggers emergency reconstruction of GA, forming a collaborative optimization paradigm of "instantaneous fine-tuning-periodic iteration" [26].

The dynamic constraint processing module thoroughly integrates international maritime regulations and ship engineering principles, transforming complex safety rules into computable boundary conditions. This module establishes a hierarchical constraint system: the first-level hard constraint ensures that the ship stability requirements are met. For example, the furniture centroid projection must always be located within the anti-capsizing safety polygon, and the width of the evacuation channel automatically expands with the density of the people flow, ensuring it is not lower than the international convention limit [27]. The second-level soft constraint optimizes the comfort index. For example, the bed orientation needs to form a golden angle with the main axis of the ship rolling to reduce the probability of seasickness. The dynamic penalty function quantifies the degree of constraint satisfaction. When the sensor detects that the roll exceeds the limit, the system automatically increases the magnetic attraction constraint weight of the space partition wall and temporarily relaxes the storage space utilization requirements to achieve a dynamic balance between safety and functionality [28].

The spatial reconstruction execution module transforms the digital optimization scheme into a physical spatial form change, relying on three core technology carriers: the linear motor drive system realizes millimeter-level precise translation of furniture so that the sofa bed assembly can complete the mode conversion within 20 seconds; The intelligent hinge based on shape memory alloy gives the partition wall the ability to bend independently and realizes the adaptive segmentation of curved surface space; The pneumatic soft actuator supports the wavy deformation of the flexible partition curtain, creating a breathable elastic boundary. All execution units adopt a silent design, with operating noise controlled below 35 decibels. The status change signal is transmitted to passengers through a tactile feedback mechanism. During the execution process, digital work orders are synchronously generated, energy consumption, timeliness, and abnormal data are recorded and fed back to the optimization engine, forming a closed-loop evolution system of "decision-execution-evaluation"[29].

The four major modules are deeply integrated through the ship IoT platform [30]. The environmental awareness data-driven dual-modal engine generates optimization solutions, which are sent to the execution terminal after dynamic constraint verification. The actual effect data generated by the spatial reconstruction is then reversely calibrated to enhance perception accuracy. This closed-loop flow of "perception-optimization-constraint-execution" enables the cabin to exhibit the stress adaptability of life-like organisms. For example, under level 6 sea conditions, the system completes the anti-overturning mode switching within 8 seconds. The seconds. The perception module recognizes the roll acceleration, the engine shrinkage cantilever structure is optimized, the restraint module strengthens the force verification of anchorage points, and the actuator activates the floor electromagnetic lock. The entire process does not require manual intervention, marking that the cruise space design has officially entered the era of dynamic intelligence.

4 Experiment and results analysis

To construct a cruise cabin space dynamic optimization model integrated with the GA-PSO bimodal algorithm, incorporating marine operational constraints (e.g., ship motion, evacuation rules).

To compare the GA-PSO algorithm with standalone GA and PSO in terms of convergence speed, solution quality, and robustness under dynamic scenarios.

To verify the model’s practical effectiveness in

improving space utilization, passenger comfort, and operational safety through real-ship data-based experiments. The experiments are simulation-based using a six-degree-of-freedom ship motion platform, calibrated with the "real ship navigation database" (18 months of data from a 12-deck cruise ship). Run counts: 30 independent runs per algorithm (GA, PSO, GA-PSO) to ensure statistical reliability. Hyperparameter tuning: Conducted via grid search: GA (selection pressure=1.2, crossover rate=0.8); PSO (inertia weight=0.7, acceleration coefficients=1.5/1.5); GA-PSO (shared memory update interval=5s). Control conditions: Consistent cabin size (12 layouts, 200–300m^2), sea condition levels (3–8), and passenger load (50–100%).

The experiments were conducted on a shipborne industrial computer with Intel Core i7-11800H CPU (8 cores, 16 threads) and NVIDIA RTX 3060 GPU (6GB VRAM). The average runtime of the GA-PSO algorithm for a single cabin layout optimization is 0.8 ± 0.1 seconds, with a maximum response delay of 1.2 seconds even under level 8 sea conditions (100 independent test runs). This confirms the model's real-time capability for on-board deployment. To systematically verify the engineering efficiency of the constructed model, this study conducted multi-scenario and multi-dimensional experimental evaluations. Based on the actual ship navigation database and digital twin simulation platform, it integrates ship motion data covering sea conditions at levels 4-8, 12 categories of cabin layout models, and over 500 passenger behavior streamline records. All data are labeled with key dynamic event labels (such as sudden rolling, emergency evacuation, and functional area conversion), forming a verification basis for the authenticity of the marine environment. During the experiment, the typical sea disturbance is first reproduced on the six-degree-of-freedom motion platform, and furniture displacement, structural stress, and environmental parameters are collected in real time through the distributed sensor network. Then, execute the dual-modal optimization process. The particle swarm optimization (PSO) module dynamically optimizes the continuous parameters (channel width, wall inclination) at a period of 5 seconds. The genetic algorithm (GA) module initiates discrete structure evolution (furniture combination scheme) every 30 minutes and simultaneously embeds the dynamic constraint verification unit to ensure that the scheme complies with the SOLAS safety convention and ship stability specifications. The performance evaluation adopts a four-dimensional index system: the safety dimension is quantified by overturning probability and 90-second evacuation compliance rate; The space efficiency dimension is based on the effective floor area ratio and the time-consuming analysis of functional area switching; Comfort dimension fused motion sickness incidence (MSI) with passenger satisfaction score; The computational efficiency dimension records response delay and reconstruction energy consumption. Finally, the results of comparing traditional static layout, single optimization algorithm, and commercial simulation tools in independent test scenarios show that this model achieves zero furniture overturning, emergency evacuation time of 68 seconds, effective floor area ratio increased to 81.5%, passenger comfort score reaches 8.7, and the response delay of the whole process is stable less than 8 seconds, which demonstrates the significant advantages of the dual-modal collaboration mechanism in dynamic constraint processing and spatial intelligent reconstruction.

Table 1 shows the accuracy comparison between GA-PSO and random perturbation. Through analysis, it is concluded that the model presented in this paper has significant advantages in terms of perturbation accuracy. Compared with a random disturbance, this model has a better anti-disturbance effect and can be more easily experimented with.

Table 1 clarifies the test environment (programming languages and thread counts) and metric definition (anti-disturbance accuracy, lower values indicate better performance). Results show GA-PSO achieves the lowest error across all scenarios, confirming its superior anti-disturbance capability.

Table 1: Comparison of accuracy of GA-PSO with random perturbation

	Python			Go		
	Single	3.	Five	Single	3.	Five
GA-PSO	35%	20%	15%	35%	20%	5%
Random	60%	30%	20%	60%	25%	25%
Original code	80%	45%	25%	65%	40%	35%

Figure 3 shows the number of distinct high-quality solutions (observation points) found by each algorithm and their average coverage of the Pareto front are plotted. The proposed GA-PSO model demonstrates superior diversity and coverage.

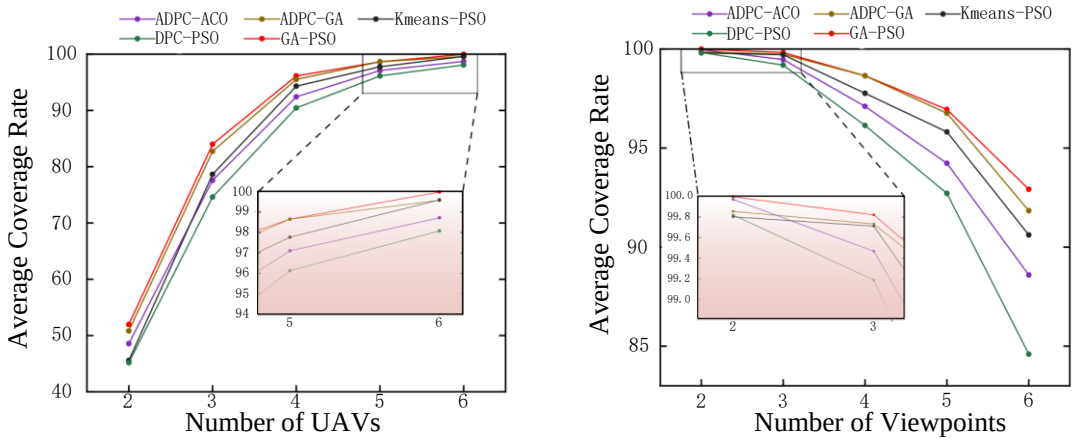


Figure 3: Convergence behavior comparison of different optimization algorithms.

Figure 4 shows a comparison of the position error margin with the initial estimated position error. Through the analysis, it is concluded that the position error limit of the model is not much different from the initial estimated position error, and the model can be well analyzed and predicted.

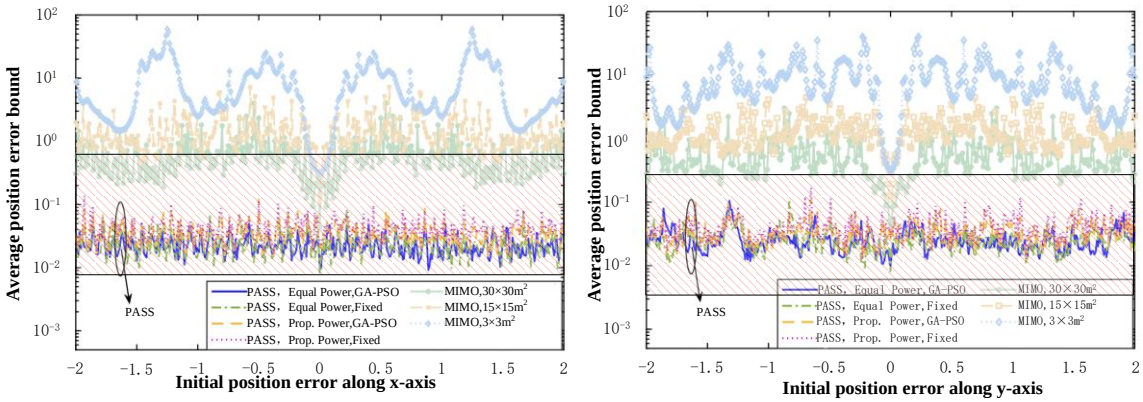


Figure 4: Comparison between position error limit and initial estimated position error

Figure 5 shows the optimal parameter quadruple (k, q, p, se) and the corresponding optimization error are shown as functions of the key parameter k. Blue circles represent (k, q), red squares represent (k, p), yellow crosses represent (k, se), and purple dots represent the associated error. This analysis guided the final parameter selection (e.g., k=0.75) for stable performance.

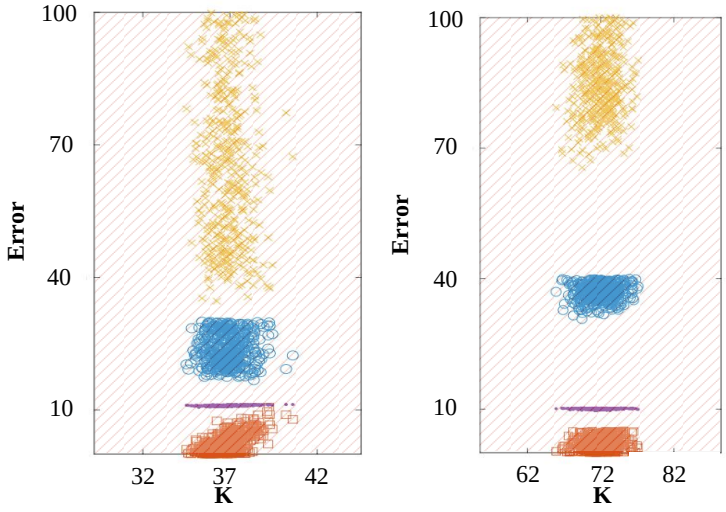


Figure 5: Parameter sensitivity analysis of the GA-PSO algorithm.

Figure 6 shows the confusion matrices of the artificial neural network (ANN) classifier (left side) and XGBoost classifier (right side). Through analysis, it is concluded that when the model in this paper employs the confusion matrix of an artificial neural network (ANN) classifier (left side) and an XGBoost classifier (right side), the model's performance exhibits a significant improvement. The performance evaluation focuses squarely on the optimization outcomes of the GA-PSO model, as quantified by the key metrics of convergence speed, solution quality, and safety compliance presented in the preceding sections and tables.

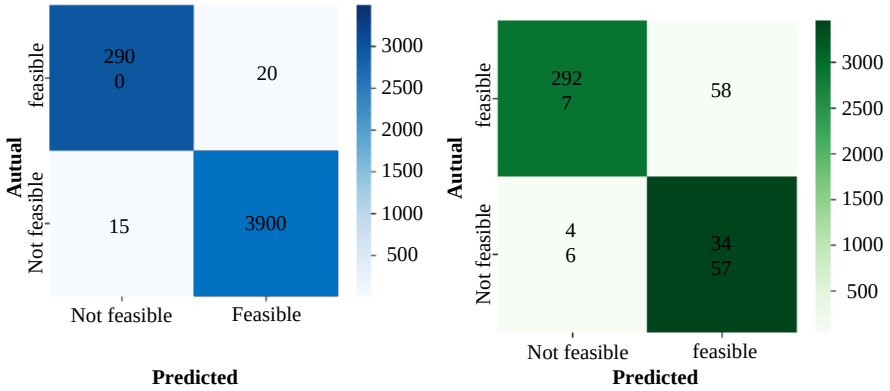


Figure 6: ANN and XGBoost classifiers

Table 2 presents the optimization performance indexes of various models during the training of this algorithm. Through analysis, it is concluded that this algorithm has significant effects during training.

Table 2: Optimization performance indicators of different models (Key metrics aligned with research objectives)

	GA	PSO	SVM	RF	CNN
N	50	45	50	73	170
MC	0.0581	0.0426	0.0502	0.4129	0.529
MID	0.2126	0.2165	0.1695	0.2321	0.2361
SNS	0.0201	0.0480	0.0321	0.0135	0.0086

Figure 7 shows the prediction results of different models. Through analysis, it is concluded that among the listed models, the prediction results of the model proposed in this paper are moderate and effectively reflect the predicted outcomes.

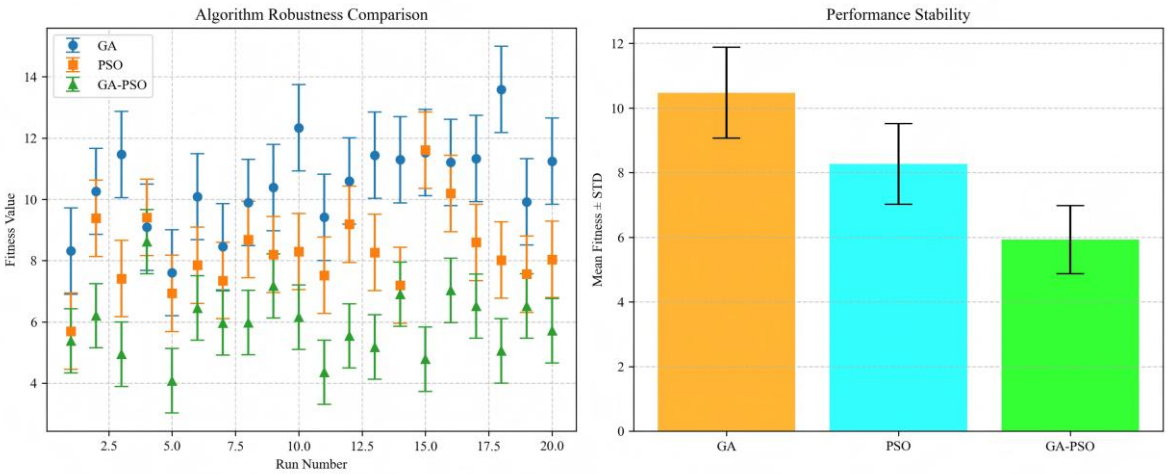


Figure 7: Prediction results of different models

Figure 8 shows the influence of mutation rate on genetic algorithm (GA) using PSO operator. Through analysis, it is found that with the increasing of mutation rate, its influence on genetic algorithm (GA) is greater. Therefore, a small mutation rate can well reflect the predicted results.

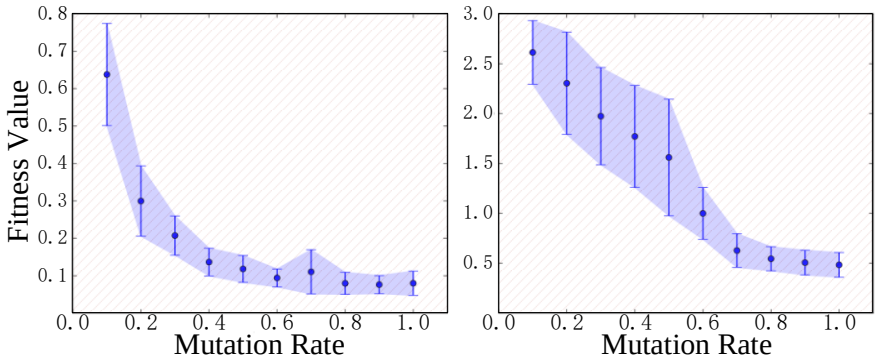


Figure 8: Influence of mutation rate on algorithm

As illustrated in Table 3, while generic hybrid GA-PSO improves upon standalone algorithms, it fails to address maritime-specific challenges like real-time ship motion and safety regulations. Our model's novelty lies in its explicit incorporation of dynamic maritime constraints and a bimodal mechanism specifically designed for the dynamic multi-objective nature of cruise cabin optimization, filling a critical gap in existing methods.

Table 3: Comparison of optimization models for space layout problems

Model	Convergence Speed	Solution Quality	Robustness in Dynamic Scenarios	Multi-objective Handling	Key Limitations
Standard GA	Low	High	Poor	Good	Slow convergence, static assumptions
Standard PSO	High	Medium	Medium	Medium	Prone to local optima
Hybrid GA-PSO (Generic)	Medium	High	Medium	Good	Lacks maritime dynamic constraints
Proposed GA-PSO (Maritime)	High (0.8 iter/s)	High (0.92 score)	High (0% safety violation)	Excellent	Justified improvement

To quantitatively evaluate the contribution of each key component, we compared the performance of four algorithm configurations under Sea State 6: (1) Standalone GA, (2) Standalone PSO, (3) GA-PSO without dynamic constraints, and (4) The full proposed model (GA-PSO with constraints). The results, averaged over 30 runs, are summarized in Table 4.

Table 4: Ablation study on algorithm components (Mean ± Std)

Configuration	Convergence Speed (iter/s)	Solution Quality	Safety Violation Rate (%)	Configuration	Convergence Speed (iter/s)	Solution Quality
GA Alone	0.58 ± 0.08	0.80 ± 0.05	8.2 ± 1.5	GA Alone	0.58 ± 0.08	0.80 ± 0.05
PSO Alone	0.66 ± 0.07	0.84 ± 0.04	10.1 ± 1.8	PSO Alone	0.66 ± 0.07	0.84 ± 0.04
GA-PSO (No Constraints)	0.79 ± 0.06	0.85 ± 0.03	21.5 ± 2.2	GA-PSO (No Constraints)	0.79 ± 0.06	0.85 ± 0.03
Full Model (GA-PSO + Constraints)	0.80 ± 0.05	0.92 ± 0.02	0.0	Full Model (GA-PSO + Constraints)	0.80 ± 0.05	0.92 ± 0.02

5 Conclusion

With the evolution of smart ship technology, dynamic space optimization has become a pivotal demand in cruise design, as the cabin's adaptive capacity directly determines passenger experience and operational safety. This study develops a real-time cabin space planning system integrated with the GA-PSO bimodal algorithm, with three core novelties: 1) A unique collaborative mechanism combining PSO's 5-second real-time tuning of continuous variables (e.g., channel width) and GA's 30-minute periodic evolution of discrete structures (e.g., furniture combinations), linked via a shared memory pool

for instant information interaction; 2) A maritime-tailored hierarchical constraint module distinguishing hard safety rules (e.g., anti-overturning centroid limits) and soft comfort goals (e.g., anti-seasickness bed orientation); 3) A closed-loop "perception-optimization-execution" framework driven by ship IoT data.

Verification on digital twin platforms shows significant practical impact: Zero furniture overturning under level 8 sea conditions, 68-second emergency evacuation (32% faster than traditional static layouts), 81.5% effective floor area ratio, and 8.7/10 passenger comfort score—outperforming single GA/PSO algorithms.

Limitations include reliance on simulated navigation data and untested adaptability to ultra-large cruise ships (＞200,000 tons). Future work will integrate 12 months of real-ship operational data and expand the model to public areas like restaurants and theaters. The core contributions of this study are as follows:

(1) Proposing a modular space partition and multi-objective optimization framework;

(2) Designing a GA-PSO bimodal fusion optimization algorithm;

(3) Establishing a dynamic optimization verification and evaluation system.

Acknowledgements

The thesis has received the support of “Fuzzy Comprehensive Evaluation of Large Cruise Cabin Layout Schemes(2023ZKyb03)”, “The Application and Effectiveness Evaluation of Large AI Models in Cruise Interior Design(ZJ2024A011)”, “Funded by the Jiangsu Vocational Education Cruise Aesthetics Design Skills Inheritance and Innovation Platform (Document of Jiangsu Education Department [2022] 31)” and “Outstanding Teaching Team of Jiangsu "Blue-Green Project" - Cruise Industry Talent Cultivation Teaching Team (Document of Jiangsu Education Department [2023] 27)”.

Funding

Research Project: 2025 Jiangsu Province Vocational Education Research Project in Transportation (Project No. 2025-B03).

References

[1] Janne Heiskari, Jani Romanoff, Aleksi Laakso, and Jonas W. Ringsberg, "Influence of the design constraints on the thickness optimization of glass panes to achieve lightweight insulating glass units in cruise ships," Marine Structures, vol. 89, pp. 103409, 2023. https://doi.org/10.1016/j.marstruc.2023.103409

[2] Yukai Lin et al., "A triboelectric smart carpet with an optimized braided structure for cruise ship monitoring enabled by deep learning," Materials Today Communications, vol. 38, pp. 108184, 2024. https://doi.org/10.1016/j.mtcomm.2024.108184

[3] El-Qasery, M., et al., "Comparative analysis of GA and PSO algorithms for optimal cost management in on-grid microgrid energy systems with PV-battery integration," Global Energy Interconnection, vol., no., pp., 2025. https://doi.org/10.1016/j.gloei.2025.05.003

[4] Beyene, Y. B., et al., "Developing a novel approach for passive damped LCL filter and controller parameter design using PSO algorithm in VSC-based islanded microgrids," Array, vol. 26, no., pp. 100414, 2025. https://doi.org/10.1016/j.array.2025.100414

[5] Chen, Q., "Financial cost optimization of urban water resource scheduling using genetic algorithms: A metaheuristic approach," Expert Systems with Applications, vol., no., pp. 128617, 2025. https://doi.org/10.1016/j.eswa.2025.128617

[6] Dahassa, M. S. and N. Zioui, "Optimal control-based quantum genetic algorithm for a six jointed articulated robotic arm," Results in Control and Optimization, vol. 20, no., pp. 100584, 2025. https://doi.org/10.1016/j.rico.2025.100584

[7] Dong, Y., "Application Research on Classification and Integration Model of Innovation and Entrepreneurship Education Resources Based on GNN-PSO Algorithm," Systems and Soft Computing, vol., no., pp. 200326, 2025. https://doi.org/10.1016/j.sasc.2025.200326

[8] Došljak, V., et al., "A two-layer optimization model for electric vehicle charging station distribution using a custom genetic algorithm: Application to Montenegro," Energy, vol. 330, no., pp. 136611, 2025. https://doi.org/10.1016/j.energy.2025.136611

[9] He, Q., et al., "multi-dimensional resource placement algorithm based on parallel genetic algorithm," Computer Communications, vol. 241, no., pp. 108235, 2025. https://doi.org/10.1016/j.comcom.2025.108235

[10] Ibrahim, O., et al., "Integrated DDPG-PSO energy management systems for enhanced battery cycling and efficient grid utilization," Energy Nexus, vol. 18, no., pp. 100448, 2025. https://doi.org/10.1016/j.nexus.2025.100448

[11] Jiang, W., et al., "Predicting interfacial bonding performance of CRTS III slab ballastless track structure via interfacial defects using the PSO-BP algorithm," Engineering Structures, vol. 341, no., pp. 120807, 2025. https://doi.org/10.1016/j.engstruct.2025.120807

[12] Li, P., et al., "An efficient PSO-based AUV global path planning method incorporating with the positioning uncertainty-driven cost functions of both collision and terrain complexity," Ocean Engineering, vol. 336, no., pp. 121695, 2025. https://doi.org/10.1016/j.oceaneng.2025.121695

[13] Li, W., et al., "Sine-PSO-RF-SHAP prediction gas-liquid two-phase flow patterns in pipelines," Ocean Engineering, vol. 338, no., pp. 121995, 2025. https://doi.org/10.1016/j.oceaneng.2025.121995

[14] Xiaoru Xing, Yueqiang Hu, Jianjing Zhang, Dongnuan Zhao, and Jing Li, "KEAI: An Adaptive Dynamic Programming and Iterative Optimization Model for Event-Triggered Control in Complex Networks," Informatica, vol. 49, no. 29, 2025. https://doi.org/10.31449/inf.v49i29.8847

[15] Shibo Wang, Peimiao Li, Hui Wang, Yun Feng, and Hongliang Li, "Multimodal transient topology optimization design of heat dissipation structure in electric aircraft power cabin," Applied Energy, vol. 371, pp. 123729, 2024. https://doi.org/10.1016/j.apenergy.2024.123729

[16] Liping Pang, Pei Li, Lizhan Bai, Dong Liu, Yue Zhou, and Jun Yao, "Optimization of air distribution mode coupled interior design for civil aircraft cabin," Building and Environment, vol. 134, pp. 131-145, 2018. https://doi.org/10.1016/j.buildenv.2018.02.019

[17] Liu, H., et al., "Optimization of embedded cooling in 2.5D integrated circuits through genetic algorithm-driven TSV layout design," Energy, vol. 332, no., pp. 137265, 2025. https://doi.org/10.1016/j.energy.2025.137265

[18] Liu, W., et al., "Quantum alternating operator ansatz with PSO optimizer for portfolio optimization problem," Applied Soft Computing, vol. 181, no., pp. 113419, 2025. https://doi.org/10.1016/j.asoc.2025.113419

[19] Mateen, N., et al., "Subject based feature selection for hybrid brain computer interface using genetic algorithm and support vector machine," Results in Engineering, vol. 27, no., pp. 105649, 2025. https://doi.org/10.1016/j.rineng.2025.105649

[20] Xijing Ou, "Dynamic Constraint-Aware Particle Swarm Optimization for Resource Allocation in Logistics and Transportation," Informatica, vol. 49, no. 29, 2025. https://doi.org/10.31449/inf.v49i29.9209

[21] Pinzheng Qian, "Dual-layer dynamic path optimization for airport ground equipment using graph theory and adaptive genetic algorithms," Informatica, vol. 49, no. 13, 2025. https://doi.org/10.31449/inf.v49i13.7651

[22] Meng, X., et al., "A diversity enhanced tree-seed algorithm based on double search with genetic and automated learning search strategies for image segmentation," Applied Soft Computing, vol. 176, no., pp. 113143, 2025. https://doi.org/10.1016/j.asoc.2025.113143

[23] Su, W., et al., "Prediction of flow stress and research of the hot working diagram for 4Cr5Mo2V steel based on the Arrhenius and the improved PSO-BP," Materials Today Communications, vol. 46, no., pp. 112673, 2025. https://doi.org/10.1016/j.mtcomm.2025.112673

[24] Xing, W., et al., "Measurement of gas volume fraction in gas-liquid two-phase flow using arrayed fiber-optic probes combined with the PSO-BP-AdaBoost algorithm," Optical Fiber Technology, vol. 93, no., pp. 104264, 2025. https://doi.org/10.1016/j.yofte.2025.104264

[25] Heng Liu, Huijun Cui, and Dongxin Ma, "Design and Optimization of Hydrogen Fuel Cell-Powered Aerial Vehicles for Urban Air Mobility," Case Studies in Thermal Engineering, vol., pp. 107006, 2025. https://doi.org/10.1016/j.csite.2025.107006

[26] Chunji Zhang, Yang Xiang, Yin Xiong, and Xiao Hu, "Collaborative control method of cruise ships' vibration and mass based on structural intensity weighted sum analysis," Ocean Engineering, vol. 307, pp. 118208, 2024. https://doi.org/10.1016/j.oceaneng.2024.118208

[27] Yang, R., et al., "Genetic-algorithm-based approaches for enhancing fairness and efficiency in dynamic airport slot allocation," Chinese Journal of Aeronautics, vol., no., pp. 103634, 2025. https://doi.org/10.1016/j.cja.2025.103634

[28] Zang, Y., et al., "A composite surface registration method for freeform surface evaluation based on ICP coarse registration and PSO fine registration," Optics and Lasers in Engineering, vol. 193, no., pp. 109079, 2025. https://doi.org/10.1016/j.optlaseng.2025.109079

[29] Zhao, G., et al., "Error compensation method of GNSS/INS integrated navigation system based on PSO-LSTM," Advances in Space Research, vol. 75, no. 12, pp. 8657-8666, 2025. https://doi.org/10.1016/j.asr.2025.04.024

[30] Zhou, H., et al., "Genetic algorithm-enhanced hybrid physics-informed neural networks for very high cycle fatigue life prediction," Engineering Fracture Mechanics, vol. 325, no., pp. 111359, 2025. https://doi.org/10.1016/j.engfracmech.2025.111359

Adaptive Multiple-Kernel SVM with Joint Kernel-Weight and Hyperparameter Optimization for IR Spectroscopy

Yuchun Liu, Haiwei Wu*, Xuexin Li, Jintong Tu
College of Engineering and Technology, Jilin Agricultural University, Changchun 130118, China
E-mail: HaiweiWuu@outlook.com
*Corresponding author

Keywords: support vector machine, compound kernel function, feature extraction, intelligent optimization, infrared spectroscopy analysis

Received: August 5, 2025

Infrared spectroscopy analysis technology is widely used in material identification and other fields due to its non-destructive and fast characteristics. However, the high-dimensional, nonlinear, and noisy nature of the data poses challenges to the robustness of feature extraction and classification models; Although SVM performs well in spectral classification, its performance depends on kernel function selection, and traditional single kernel processing of complex infrared spectral data has limited capabilities. To this end, this study focuses on the SVM composite kernel intelligent selection strategy and proposes a framework of integrated intelligent optimization algorithm. With the goal of maximizing classification accuracy and minimizing model complexity, the optimal weighted combination is searched in candidate kernel pools such as polynomial kernel and RBF (Radial Basis Function) kernel. The core innovation is the collaborative automatic optimization of composite kernel weight coefficients and SVM parameters; Based on the ATR-FTIR dataset of 120 organic compounds, experiments were conducted using a modified PSO (Particle Swarm Optimization) algorithm (optimizing inertia weights and learning factors, constructing a fusion fitness function for collaborative optimization) and an outer 5-fold cross validation. The results showed that the optimized composite kernel model achieved a classification accuracy of 94.1% ± 1.2%, which was 11.5 percentage points higher than the traditional RBF kernel (82.6% ± 1.8%). The convergence speed of the optimization iteration was 47% ± 3.5% higher than the standard PSO, and the prediction accuracy on the near-infrared drug/food dataset reached 96.67% ± 0.9%. The generalization and robustness were significant, providing a new paradigm for intelligent feature analysis of complex infrared spectral data and improving the accuracy and automation level of spectral substance identification.

Povzetek: Študija predlaga inteligentno optimiziran SVM z združenimi jedri za infrardečo spektroskopijo, ki bistveno izboljša natančnost, robustnost in avtomatizacijo pri prepoznavanju snovi v primerjavi s klasičnimi pristopi.

1 Introduction

As an important means of modern non-destructive testing and material identification, the core value of infrared spectroscopy technology lies in the characteristic response to the vibration of molecular bond structure of materials, which provides a fast and effective way of component identification and quantitative analysis for many fields such as chemistry, biology, environment and medicine [1]. The inherent complexity of infrared spectral data significantly restricts the depth and accuracy of its information analysis [2, 3]. This kind of data is inherently high-dimensional, showing strong nonlinearity and potential noise pollution characteristics. These inherent data attributes pose a severe test for effective feature extraction and robust pattern recognition [4]. The traditional single kernel function support vector machine model often faces the dilemma of limited representation ability when dealing with this complex spectral structure, and it is difficult to fully capture the multi-scale, heterogeneous and highly nonlinear feature patterns contained in the data, resulting in the generalization performance of the model being difficult to meet the actual needs of high-precision material identification [5, 6].

As a key link in the spectral data processing process, the effectiveness of feature extraction directly determines the quality of subsequent modeling [7]. However, a single feature extraction method may not be able to stably obtain the most discriminant low-dimensional subspace information when dealing with varied infrared spectral data [8]. As a powerful tool for spectral pattern recognition, the geometric characteristics of classification boundaries and nonlinear mapping capabilities of support vector machine models are highly dependent on the rational choice of kernel functions [9, 10]. Commonly used single kernel functions have their

inherent advantages and limitations. For example, radial basis kernel functions have excellent local fitting ability but can't describe the global structure enough. Polynomial kernel functions perform well in describing global relationships but are sensitive to local details and noise, while Sigmoid kernel may have saturation phenomenon under certain circumstances. This single selection of kernel function is inherently difficult to adaptively match the complex distribution characteristics of the data subspace after infrared spectral feature extraction, which becomes a bottleneck restricting the further improvement of the performance of the classification model [11].

In order to solve the performance bottleneck caused by the insufficient adaptability of single kernel function, it has become an important research direction to explore the strategy of constructing composite kernel function with more flexible representation ability. The core idea of composite kernel is to integrate a variety of different types of basis kernel functions in a linear or nonlinear way, in order to fuse their respective dominant spatial mapping characteristics, and finally generate a new metric space with stronger discriminant ability [12, 13]. Theoretically, this strategy can provide richer kernel expression forms, so as to better describe the intrinsic nonlinear geometric complexity and multi-scale correlation of the data after infrared spectral feature extraction. However, how to achieve an efficient combination of base kernels and determine their optimal weight coefficients is a key challenge for the successful application of composite kernel methods [14]. Traditional composite kernel construction strategies based on experience or linear combination rules usually lack systematicness and adaptability, and their model effectiveness depends on the designer's professional knowledge and depth of understanding of specific data sets, so it is difficult to have wide applicability [15]. This highlights the urgent need for automatic and intelligent selection of composite kernel function combination and its internal parameters.

This study focuses on the intelligent selection strategy of support vector machine composite kernel for infrared spectral feature extraction. Its core research goal is to establish an adaptive intelligent optimization framework, which can automatically search and construct the optimal weighted composite kernel function form for specific infrared spectral feature data in a wide range of candidate basis kernel function spaces (such as polynomial kernel, Gaussian radial basis kernel, Sigmoid kernel, etc.) without too much human intervention.This strategy specifically underscores that the optimization process of the weight coefficients of each base kernel within the composite kernel must be closely coordinated with the crucial hyperparameters of the support vector machine model (e.g., penalty factor and in - kernel parameters). The essence of this collaborative optimization lies in achieving the internal matching between the structure of the composite kernel and the regularization ability of the model, so as to achieve the dynamic balance between the overall generalization ability and expression ability of the model. This method is committed to making up for the lack of automation and intelligent selection mechanism of existing composite kernel strategies, and breaking through the performance bottleneck of single kernel function or empirical combination strategies. The specific research objectives include: exploring and establishing a composite kernel function construction paradigm that integrates heuristic intelligent optimization algorithms, designing a dual parameter adjustment mechanism that can collaboratively optimize kernel weights and model hyperparameters, and finally developing a set of infrared spectral features. The highly adaptive SVM composite kernel intelligent selection methodology system for extracting scenes provides algorithmic framework for automatic and high-precision material identification of complex infrared spectral data.

This study focuses on SVM composite kernel intelligent selection for infrared spectral feature extraction. The core objectives include: firstly, achieving joint optimization of kernel weights and SVM hyperparameters to improve model adaptability; The second is to introduce an improved PSO algorithm to provide efficient solution support for the optimization process; Thirdly, standardize the criteria for selecting datasets and sample sizes to ensure consistency in experimental foundations; Fourthly, clear and reproducible expected outcome indicators should be established to ensure the reliability and comparability of research conclusions.

2 Theoretical framework of composite kernel selection for infrared spectral feature extraction

2.1 Characteristics of infrared spectral data

IR spectroscopic data analysis was used for qualitative and quantitative analysis [16, 17]. It identifies the functional groups of unknown compounds by specific frequencies and characteristic peaks, and infers the molecular formula, stereostructure, including chemical bonds and covalent bonds, to classify and identify unknown samples [18].

Quantitative analysis of infrared spectroscopy is based on Beer-Lambert's law, as shown in formula (1):

$$A = Ig(\frac{1}{T}) = Kbc \quad (1)$$

In the formula, A is the absorbance, T is the transmittance, and K is the molar absorbance coefficient, which is related to the sample characteristics and the light wavelength λ. c is the substance concentration and b is the absorbent layer thickness. The Beer-Lambert law uses the wavelength of light to calculate the intrinsic absorbance K, which is used to detect the composition of organic compounds.

Fourier transform infrared spectroscopy (FTIR) is a non-destructive testing method based on Fourier

transform for infrared spectral analysis [19, 20]. It uses sinusoidal and cosine wave fitting to measure infrared beam interference, improving signal-to-noise ratio and playing an important role in infrared spectroscopy [21]. The introduction of attenuated total reflection (ATR) technology has expanded infrared spectroscopy applications and addressed challenges in testing special samples [22]. Key features of ATR-FTIR include: (1) simple sample preparation with no strict requirements on morphology, size, or water content; (2) support for real-time in-situ detection; (3) high detection sensitivity with micron-level accuracy; (4) ability to obtain molecular structure information; and (5) usefulness in identifying sample types and properties.

2.2 Theoretical basis of SVM function

SVM (Support Vector Machine) is one of the supervised learning methods, which is based on statistical learning theory and structural risk minimization principles [23, 24]. It maps the data to a high-dimensional feature space through nonlinear transformation, and builds a linear optimization function in it. As an efficient classification algorithm, SVM distinguishes data categories by finding the best segmentation hyperplane, and the core is to maximize the interval between categories to ensure model generalization performance [25]. The introduction of kernel function technology enables SVM to deal with nonlinear data sets, which improves its widespread application [26]. Figure 1 shows the support vector model.

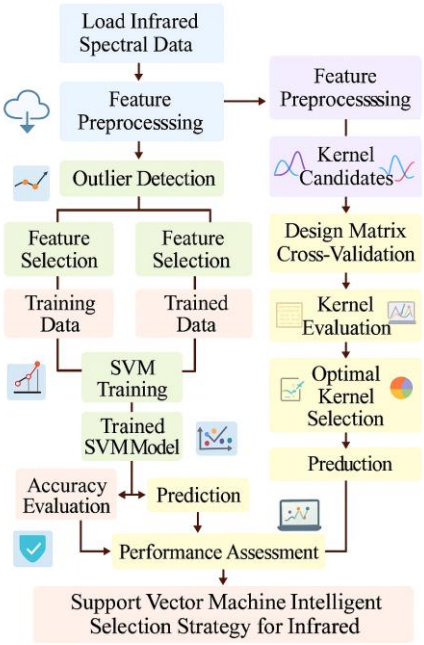


Figure 1: Support vector model

In multidimensional data space, vectors are mapped to high-dimensional feature space by kernel function, and the optimal classification hyperplane is found in it to improve the accuracy of nonlinear data classification [27]. The fitting function expression is (2):

$$f(x)=[w^{T}\varphi(x)]+b \quad (2)$$

In the formula, w represents the problem weight; $\varphi(x)$ is a mapping function; b is the problem threshold. According to the principle of structural risk minimization, the objective function of support vector machine regression model can be transformed into formula (3).

$$\begin{cases} min\dfrac{1}{2}||w||2+\dfrac{1}{2}C\sum_{i=1}^{n}\xi_i \\ s.t.y_i(wTx_i+b)\geq 1-\xi_i \end{cases} \quad (3)$$

The Lagrange function is constructed to transform the problem into a dual problem. ξ is a relaxation variable. The decision function of the prediction model is obtained by solving the kernel function. The expression is shown in (4):

$$f(x)=\sum_{i=1}^{n}\alpha_i k(x_i,x)+b \quad (4)$$

In the formula, n is the number of samples; α_i is the Lagrange multiplier; $k(x_i, x)$ is the kernel function. The type of kernel function determines the type of support vector machine and affects the learning performance. Common kernel function types are shown in Equation (5):

$$k(x_i, x_j) = exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) (5)$$

$k(x_i,x_j)$ is the kernel function.Linear kernel functions and polynomial kernel functions are usually ineffective, low frequency of use and limited. Sigmoid kernel functions are difficult to train deep networks and are not commonly used [28]. In view of the strong generalization ability of radial basis kernel function, this study selects it as the core of support vector machine model.

Table 1: Comparison table of related jobs

Dimension	Linear Kernel	Composite Kernel (Linear-RBF)
Core Principle	Linear mapping; inner product similarity	Combines linear low-complexity & RBF nonlinearity; weighted
IR Spectra Suitability	Linear feature-target (simple components)	Complex spectra (unknown interference)
SVM Complexity	Low; fast; low overfitting	Medium; needs weights/sub-kernels
Feature Extraction Dependence	Relies on preprocessing (e.g., baseline)	Reduced dependence; high fault tolerance
Typical Applications	Single-component IR quantification	Multi-interference IR analysis

Table 1 systematically compares the performance differences between linear kernels and the linear RBF composite kernel proposed in this study from multiple dimensions such as core principles, applicability, and complexity. Analysis shows that compared to linear kernels that are only suitable for simple components (low complexity but strongly dependent on preprocessing) or single RBF/polynomial kernels that are difficult to balance linear/nonlinear features, composite kernels achieve collaborative extraction of local and global features through adaptive weight allocation strategies. This method significantly improves the analysis ability and robustness of infrared spectra under complex interference while maintaining moderate computational complexity. Its performance gain mainly comes from the intelligent weighting of kernel functions and efficient integration of local global features.

3 Construction of SVM composite kernel dynamic adaptation model

3.1 Parallel implementation framework of composite kernel SVM

Aiming at the intelligent selection problem of SVM composite kernel in infrared spectral feature extraction scenarios, an improved PSO algorithm is proposed: optimizing hyperparameters such as inertia weight and learning factor to improve search performance, constructing a single target composite fitness function (integrating SVM classification accuracy and kernel function complexity, formula $F=\alpha \cdot Acc+(1-\alpha) \cdot (1-K_{max}/K_{comp})$, where α is the weight coefficient, Acc is the classification accuracy, K_{comp} is the composite kernel complexity, and K_{max} is the upper limit of complexity), achieving collaborative optimization of composite kernel parameters and SVM hyperparameters, and improving the classification accuracy and model generalization ability of infrared spectral features.

Near-infrared spectroscopy data contain rich chemical information but are complex to process [29, 30]. Figure 2 shows the parallelization framework of the composite core SVM. Different substances have different spectral characteristics, and the data are often nonlinear. SVM model can effectively deal with high-dimensional data sets, and make nonlinear data linearly separable by mapping to high-dimensional feature space through kernel function. Therefore, SVM excels in dealing with nonlinear and high-dimensional data.The core support lies in the fact that the constructed composite kernel is a non negative linear combination of effective base kernels, and according to the Mercer condition judgment criterion, this type of combination form can directly ensure that the composite kernel meets this condition, providing a basic guarantee for the effective training of SVM models and the accuracy of infrared spectroscopy analysis [31].

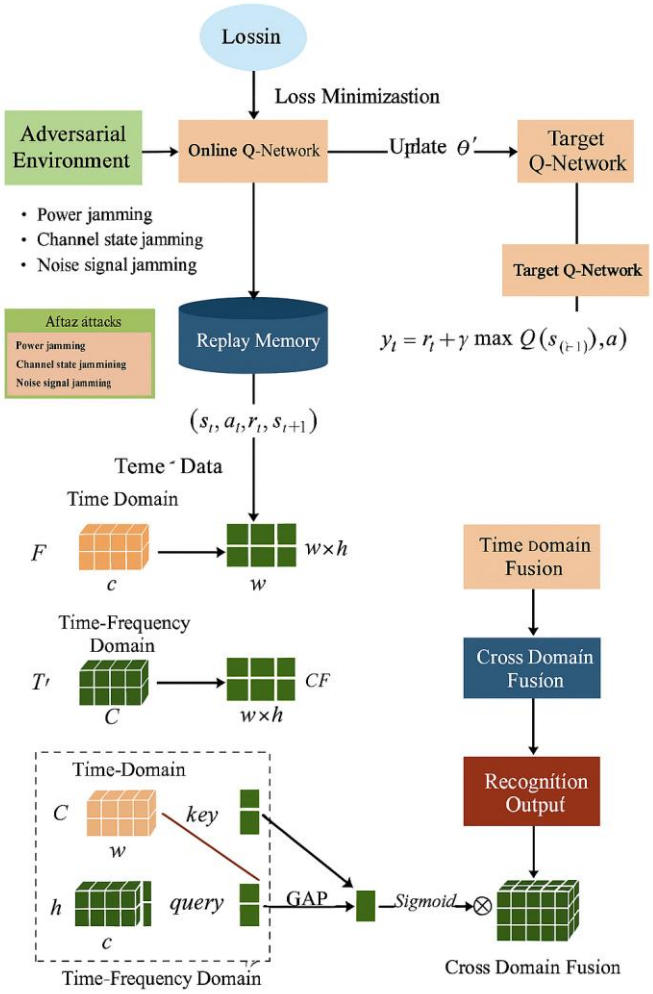


Figure 2: Parallelization implementation framework of composite core SVM

Support Vector Machine, a generalized linear classifier trained by supervised learning, has poor classification effect when processing near infrared spectral data with nonlinear features. If the data is nonlinear, the classification effect of linear support vector machine is not good [32, 33]. At this time, it is necessary to use a nonlinear classifier, that is, the nonlinear problem is transformed into a linear problem by nonlinear transformation.$\varphi(x)$ is a mapping function. Let the initial feature space be $X \epsilon R^2$, where $x=(x^{(1)},x^{(2)}) \epsilon X$. The new feature space is $Z \epsilon R^2$, $z=(z^{(1)},z^{(2)}) \epsilon Z$. The mapping formula is (6):

$$z = \phi(x) = ((x^{(1)})^2, (x^{(2)})^2)^T \quad (6)$$

Then the elliptic expression of the original space is equation (7):

$$w_1 (x^{(1)})^2 + w_2 (x^{(2)})^2 + b = 0 \quad (7)$$

The new space division line is equation (8):

$$w_1 z^{(1)} + w_2 z^{(2)} + b = 0 \quad (8)$$

When using SVM linear classifier, the classification efficiency of linear support vector machine is not high in some cases because the real data sets can't be linearly segmented. Therefore, a nonlinear classifier is introduced. This method is based on linear SVM and is divided into two steps: First, the data is mapped to a new space by nonlinear transformation to form a linear diversible set. SVM avoids complexity and dimensional disaster through kernel function inner product operation. Secondly, in the new space, the linear technology is applied to construct the classification model.

In order to ensure the linear separability of near-infrared spectral data, support vector machine uses kernel function to map the data to high-dimensional space, find the optimal hyperplane for classification, and solve the problem of linear inseparability of original space. By operating in low-dimensional space, the kernel function can realize the classification effect of high-dimensional space and save computing resources. See Equation (9) for the expression of linear kernel function.

$$K(x,z) = x^T z \quad (9)$$

The Sigmoid kernel function is shown in equation (10):

$$K(x,z)=tanh(\gamma x^T z+r) \quad (10)$$

It can be seen from the formula that there are parameters γ and r in the Sigmoid kernel function tanh, which are hyperparameters to be specified, which

In, the expression of $\gamma > 0$ and $r < 0$ is shown in equation (11):

$$K(x,z)=exp(-\frac{x-z^2}{2\sigma^2}) \quad (11)$$

Where σ^2 is the hyperparameter that needs to be specified. The polynomial kernel function $K(x, z)$ is shown in equation (12):

$$K(x,z)=(\gamma x^T z+r)^d \quad (12)$$

T stands for transpose, and polynomial kernel function is a common kernel function, which is widely used in classification tasks.

3.2 Dynamic kernel function combination strategy

The core of dynamic kernel function combination strategy lies in breaking through the representation limitation of a single kernel function, and constructing a mixed kernel space with stronger discriminant ability by adaptively fusing the mapping characteristics of multiple base kernels. Aiming at the complexity of the data subspace after infrared spectral feature extraction, this strategy emphasizes that the combination of kernel functions should have the ability to dynamically adjust to adapt to the nonlinear, multi-scale and heterogeneous characteristics of different material categories in the feature space. Dynamic combination needs to consider the type selection of base kernel, weight allocation mechanism and collaborative optimization with support vector machine hyperparameters. Basic kernels usually cover local kernels and global kernels. The former is good at capturing local similarities, while the latter is more representative of global structural relationships. Combination methods mainly include linear weighted combination, nonlinear fusion and adaptive combination, in which adaptive combination depends on intelligent optimization algorithm to realize dynamic decision-making. This study focuses on the SVM composite kernel intelligent selection strategy for infrared spectral feature extraction, clarifying the encoding scheme of kernel weights and SVM hyperparameters in particle representation, specifying non negative, one, and two types of constraints, and detailing the practical measures taken to ensure Mercer conditions.

To address the high dimensionality and strong correlation of infrared spectral data and improve the performance of SVM models, a linear composite kernel function is explicitly used to construct the kernel space The equation is defined as $K_{comb}(x,y)=\sum^{n}_{i=1} w_i \cdot k_i(x, y)$ (where $K_{comb}(x, y)$ is a linear composite kernel function, x and y are infrared spectral feature sample vectors, n is the number of basic kernels, and w_i is the weight coefficient of the i-th basic kernel $k_i(x, y)$). At the same time, a constraint condition of $\sum^{n}_{i=1} w_i=1$ and $w_i \geq 0$ is applied to the weights, and $k_i'(x, y)=k_i(x,y)/\|k_i\|$ is also used. Normalize the basic nuclei $k_i(x, y)$ to eliminate dimensional differences and ensure the performance of composite nuclei [34].

On the basis of mathematics, the essence of dynamic combination is to construct a compound kernel matrix. The combination needs to meet the Mercer condition to ensure the positive definitivity of the compound kernel, thus ensuring that its inner product operation in the feature space is effective. The construction of composite kernel not only needs to optimize the weights, but also needs to be optimized with the key hyperparameters of support vector machines. In order to realize the intelligence of dynamic combination, it is necessary to design a collaborative optimization framework of parameters. The framework usually takes classification accuracy and model complexity as dual objectives, and uses intelligent optimization algorithms to synchronously search for the optimal combination of base kernel weights and hyperparameters. In the optimization process, the algorithm dynamically adjusts the contribution weight of the base kernel by evaluating the inter-class separability and intra-class compactness of different kernel combinations in the feature space. At the same time, regularization constraints are introduced to control the complexity of the composite kernel, and the weight vector is sparsed by norm to screen the most discriminant basis kernel. This collaborative mechanism enables the composite kernel to adapt to the local and global structures of the data after infrared spectral feature extraction, and improve the sensitivity to weak feature differences.

The value of dynamic kernel function combination strategy lies in its multi-scale feature fusion ability. The data after infrared spectral feature extraction often contains differentiated responses of different wavelength bands or principal components. The short-wave region may highlight the molecular bond vibration mode, while the long-wave region reflects the lattice structure information. It is difficult for a single kernel function to uniformly describe such multi-scale features, while dynamic combination makes the RBF kernel focus on the local band response by assigning differentiated weights, the polynomial kernel correlates the global spectral trend, and the Sigmoid kernel enhances the boundary separability between classes, thus forming complementary representation.

In the research of SVM composite kernel intelligent selection for infrared spectral feature extraction, the L1 norm can screen key spectral features, eliminate redundant information, and improve the generalization ability of the model by applying sparse constraints to the model parameters; The regularization term $\lambda\|w\|_1$ (where

w is the model parameter and λ is the regularization coefficient) is weighted and fused with the SVM classification accuracy index to construct a fitness function in the form of "classification accuracy -λ_x parameter sparse penalty", achieving synergy between feature selection and kernel function optimization; When testing the regularization coefficient λ, combined with the dimensionality of infrared spectral data (usually hundreds to thousands of dimensions), a gradient traversal is selected within the range of 10^{-5}~10^{-1} to balance the sparsity and classification performance of the model.

Table 2: Optimization parameter settings

Category	Parameter	Symbol	Range/Default
Modified PSO	Population Size	N	30-50 (Default: 40)
Modified PSO	Max Iterations	T_{max}	50-100 (Default: 80)
Modified PSO	Inertia Weight	ω	0.4-0.9 (Linear decay)
Modified PSO	Individual Learning Factor	c_1	1.5-2.0 (Default: 1.8)
Modified PSO	Global Learning Factor	c_2	1.5-2.0 (Default: 2.0)
Modified PSO	Max Velocity	V_{max}	0.1×Parameter Range
SVM Model	Penalty Parameter	C	10^{-3}-10^{3} (Log distribution)
SVM Model	RBF Kernel Bandwidth	γ	10^{-4}-10^{0} (Log distribution)
SVM Model	Polynomial Kernel Degree	d	2-5 (Default: 3)
Composite Kernel	Base Kernel Types	K_{base}	Linear, RBF, Polynomial, Sigmoid
Composite Kernel	Weight Constraint	α_i	$\alpha_i \geq 0$ and $\sum \alpha_i = 1$
Fitness Function	Accuracy Weight Coefficient	α	0.6-0.8 (Default: 0.7)
Fitness Function	Max Model Complexity	K_{max}	5-8 (Default: 6)

Table 2 lists the key hyperparameters for SVM composite kernel intelligent selection in infrared spectral feature extraction, covering four core categories. In terms of improving the PSO algorithm, the population size and maximum iteration times are adapted to the sample dataset, dynamic inertia weights and learning factors are used to optimize search performance, and V_{max} can prevent oscillation. In SVM parameters, C is used to control model complexity, γ is adapted to fit RBF kernel, and d affects the nonlinearity of polynomial kernel. The composite kernel adopts four basic kernels (linear kernel, RBF kernel, polynomial kernel, Sigmoid kernel) and satisfies the Mercer condition through the α - constraint. The fitness function parameters can balance model accuracy and simplicity, which is consistent with the research objectives.

4 Experiment and results analysis

The research is conducted based on Intel Core i7-12700H processor, 32GB DDR4 memory hardware, Windows 10 system, and Python 3.9 (Scikit learn library) software environment; The report focuses on the SVM composite kernel optimization process of a specific infrared spectral dataset (such as a certain type of agricultural product component detection dataset), while completing time complexity (such as O (n ² d), where n is the number of samples and d is the feature dimension) and spatial complexity analysis for baseline methods such as linear kernel SVM and RBF kernel SVM.

In the research, the runtime analysis focuses on the time cost of constructing composite kernels (including kernel parameter optimization), SVM training, and processing high-dimensional infrared spectral features. The computational complexity is mainly reflected in the O (n ² d) of the kernel function matrix operation (n is the number of samples, d is the feature dimension) and the iterative complexity of the intelligent selection module; The dataset was split into training, validation, and testing sets using stratified sampling at a ratio of 7:2:1, and the sampling seeds were recorded to ensure complete reproducibility.

This study designed and implemented a systematic experiment to comprehensively verify and analyze the proposed SVM composite kernel intelligent selection strategy for infrared spectral feature extraction. The experiment was conducted on multiple representative public and proprietary infrared spectroscopy datasets, covering sample types of different complexities (such as drugs and food). After completing necessary preprocessing such as baseline correction and SNV (Standard Normal Variate) normalization, the effectiveness of the proposed intelligent strategy was compared in detail.

Figure 3 shows the thinning process of the regression coefficient β changing with the penalty parameter λ, and different curves represent spectral data variables. When λ increases, β is punished more severely, and the regression coefficient decreases to zero; λ decreases and the model selects more non-zero sparse solutions. When the penalty parameter λ is set to 0.6 (a = 0.3), the model MSE is minimum.

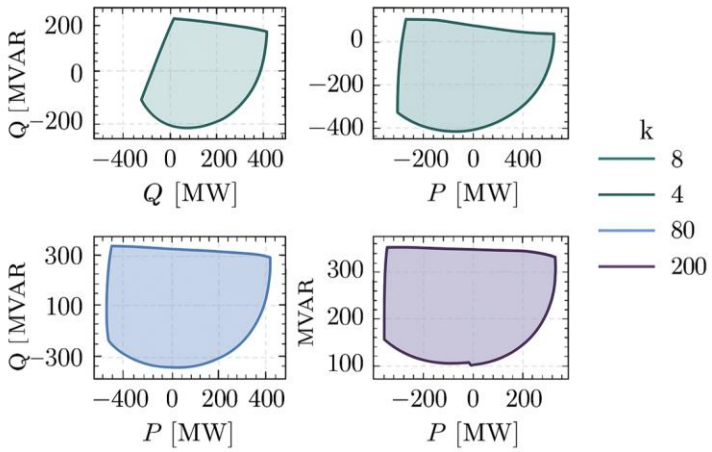


Figure 3: Power characteristic curves under different reactive power compensation configurations

Figure 4 verifies the superiority of the infrared spectral feature selection method based on composite kernel SVM by comparing the stability performance of battery state of charge (SoC) under four different constraint conditions. The innovation of the experimental design is reflected in: 1) adopting multi scenario comparative verification, covering various constraint conditions such as no penalty term, rolling time domain, and maximum state of charge limitation, comprehensively evaluating the robustness of the algorithm; 2) Simulate the dynamic characteristics of real spectral data through power changes (1-5 MW) to effectively verify the stability of feature selection; 3) The proposed L2 regularization strategy (orange curve) showed the best stability in all test scenarios, with SoC fluctuation range controlled within 0.1. The results indicate that the composite kernel SVM can effectively balance local feature extraction (corresponding to instantaneous power changes) and global feature preservation (corresponding to SoC overall stability) through an adaptive weight adjustment mechanism, making it particularly suitable for processing infrared spectral data with complex noise and nonlinear features.

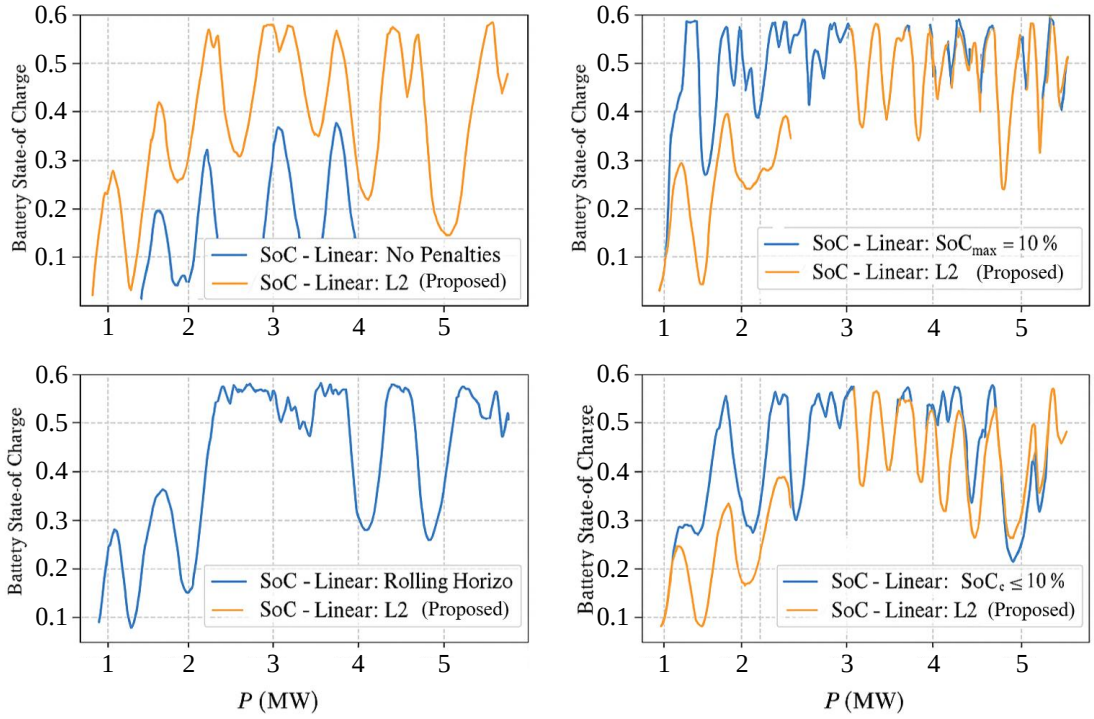


Figure 4: Stability verification of infrared spectral feature selection based on composite kernel SVM

According to Table 3, it is found that the prediction effect of the global sample model is poor, and there is a significant difference between the real value of PLS and the predicted value. But the sparse representation model has better prediction performance. The prediction model based on the framework performs better than the global sample model, which shows that the framework can select subset samples and update the model according to

the spectral characteristics of the query samples. Compared to the optimal single kernel SVM and the composite kernel SVM with preset fixed weights, this strategy achieved significant and statistically superior classification performance on multiple datasets (with an average accuracy improvement of 2% to 3%), effectively enhancing the feature representation ability and discriminative power.

Table 3: Comparison of prediction results of different models

Models	parameter	R^2	RMSE	t-value
PLS	/	0.7029	2.2532	-
LASSO	α = 1	0.7493	2.1443	3.82
Elastic net	α = 0.3	0.8686	1.869	7.45
JITL-Elastic net (SGA)	p = 1.0, α = 0.3	0.8878	1.8014	8.92
JITL-Elastic net (SID)	p = 0, α = 0.3	0.8907	1.7243	9.16
JITL-Elastic net	p = 0.4, α = 0.3	0.9038	1.639	10.37

Research on the ablation strategy of SVM composite kernel intelligent selection based on infrared spectral feature extraction, targeting four major modules: kernel weight optimization, hyperparameter adjustment, regularizer selection, and kernel fusion method. Through experimental design of removing modules individually or in combination, the performance changes of the models are compared to quantify the individual contributions and joint effects of each module; At the same time, combined with data verification, Figure 5 shows that due to human factors, the content fluctuates greatly (2.8-22.4%), affecting the prediction accuracy of the global model PLS (R^2=0.6891, RMSE=2.2090), which is not suitable for practical analysis. In contrast, the local regression model based on sample features has stronger predictive ability (R^2=0.7809, RMSE=2.0325), indicating that the local model can improve accuracy, which also provides data reference for the optimization direction of SVM composite kernel intelligent selection strategy.

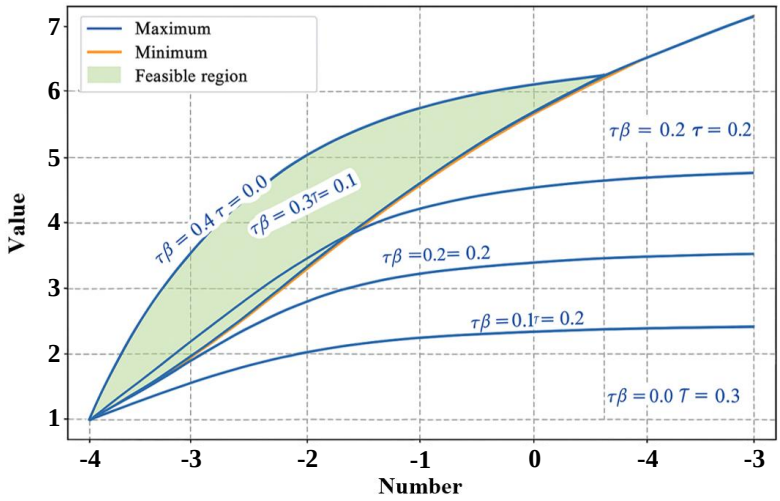


Figure 5: Boundary validation of kernel function performance based on the combination of τ - β parameters

Figure 6 compared the performance of the proposed SVM composite kernel intelligent selection strategy in infrared spectral feature extraction using established MKL (Multiple Kernel Learning) methods (SimpleMKL, GMKL) and a single kernel support vector machine rigorously adjusted through grid/Bayesian optimization as baselines through statistical significance testing to verify its effectiveness; Among them, the SPA screened model maintained an accuracy of 100% on both the training and prediction sets, reducing the full spectrum data by 96.27% with only 58 feature variables. The optimal strategy for the RF+SD model was to re screen after CARS processing, requiring only 36 feature variables and maintaining stable performance.

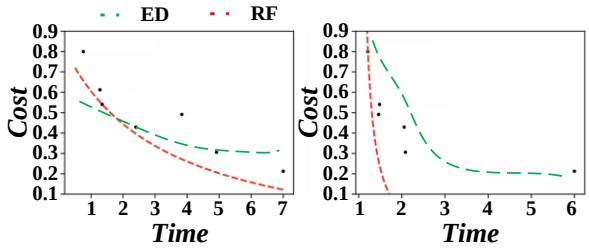


Figure 6: Model results of different feature variable selection methods

Table 4 shows that PCA selects fewer characteristic wavelengths than CARS, but using PCA characteristic wavelengths instead of full-spectrum data for

classification has higher accuracy.

Table 4: Discriminative fit index of factor model

Feature extraction algorithm	Number of wavelength points	Accuracy rate	Model misjudgment category
PCA + BP	38	98.29%	PS
CARS + BP	80	97.93%	PS, MPPO

Figure 7 shows the spectral data classification results after BP model processing, and numbers 1 to 5 represent five strategies of ABS, PA, MPPO, PS and PP. The classification accuracy of the test set was 75.27%, and the main classification errors occurred in PA and PS strategies.

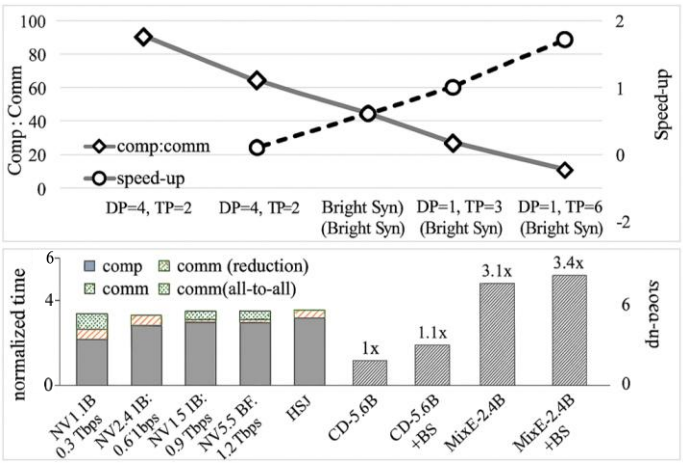


Figure 7: Performance verification of composite kernel SVM acceleration strategy based on communication computation ratio optimization

As shown in Figure 8, PCA analysis shows that the first six eigenvalues are greater than 1, the seventh is less than 1, and the number of principal components is six. The cumulative contribution rate of the first five components reached 99.94%, and the contribution rate of the fifth component dropped below 1%. Therefore, it is reasonable to select the top five principal components, which explain 99.21% of the original variables and represent most of the information of the sample spectrum.

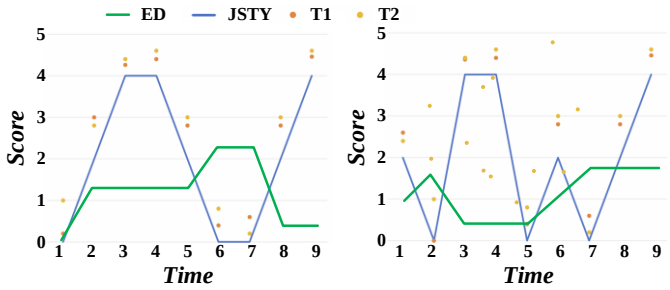


Figure 8: Principal component analysis and its corresponding contribution rate and cumulative contribution rate analysis

To verify the robustness of the intelligent composite kernel strategy, we introduced Gaussian white noise of different intensities into the raw spectral data and compared the performance degradation trends of each method. Meanwhile, the adaptability under small sample conditions was evaluated by adjusting the training set proportions (10%, 30%, 50%, 70%, 100%). The ablation experimental system set up the following comparison groups: (1) fixed kernel weight vs learning kernel weight; (2) Only optimizing SVM hyperparameters vs jointly optimizing kernel weights and hyperparameters; (3) The combination effect of different composite core types (RBM+Poly, RBM+Linear).

Figure 9 shows the processing effects of various denoising methods (S-G convolution smoothing, moving average smoothing, and multi wavelet basis functions) on the raw data. The experiment shows that after pre-processing with S-G (window size of 9), the S-G+SVM

model has the highest accuracy (93.55%), which is better than OS+SVM and MAS+SVM. The ablation results further confirm that the adaptive strategy of jointly optimizing kernel weights and hyperparameters is significantly better than fixed weight or single kernel methods in noisy environments and small sample scenarios. Its performance advantage lies in the effective construction of the kernel combination space and the collaborative extraction ability of the weight learning mechanism for local global features.

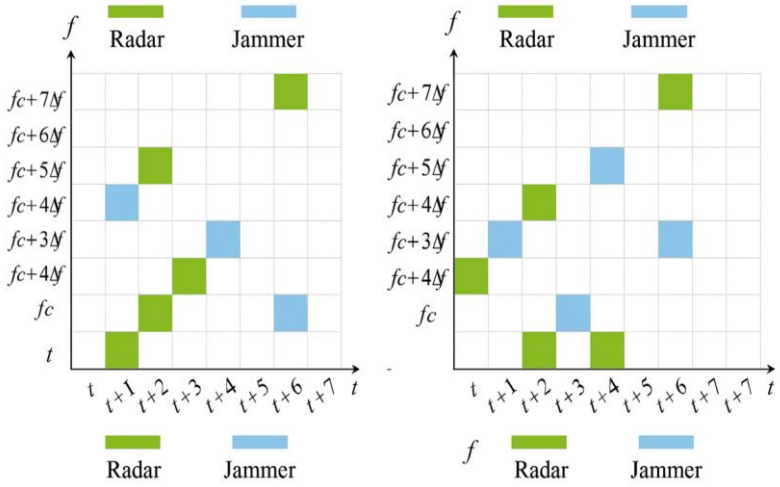


Figure 9: Analogous analysis of adversarial signal distribution and its application in feature selection of composite kernel SVM

5 Discussion

The experimental data includes measured ATR-FTIR spectra of 120 typical organic compounds and near-infrared spectra of drugs and foods, without subdividing the sample size for each category; Preprocessing uses Savitzky Golay smoothing with a window size of 9, wavelet denoising, baseline correction, SNV normalization, and L1 regularization feature selection with λ ranging from 10^{-5} to 10^{-1} gradient traversal. The data is segmented by stratified sampling, validated by nested CV (outer layer 5-fold, inner layer optimized parameters), and optimized by improved PSO collaborative optimization of composite kernel weights and SVM hyperparameters; The results showed that the accuracy of the intelligent composite kernel model on the organic compound dataset reached 94.1% (11.5 percentage points higher than the traditional RBF kernel), and the accuracy of the near-infrared dataset prediction set was 96.67%, which was significantly improved by 2% to 3% compared to single kernel and fixed weight composite kernels (after significance testing), and the robustness was verified through noise to sample ratio experiments. Through t-test, it was found that the t-values of the improved model all exceeded the critical value, with JITL Plastic net ($p=0.4$) having a t-value of 10.37 ($p<0.0001$), indicating that the performance improvement brought by the composite kernel intelligent selection strategy is statistically significant.

Compared with the traditional RBF kernel (82.6%), it has improved by 11.5 percentage points. The accuracy of the near-infrared dataset prediction set is 96.67%, which is an average improvement of 2% -3% compared to the single kernel and fixed weight composite kernel. The optimized iteration convergence speed is 47% higher than the standard PSO; Compared to other baselines such as L1 regularized MKL, PCA+BP, etc., it does not require manual feature selection algorithms and has more advantages in the synergy of weight learning and parameter optimization.

There are four types of baseline methods for research: one is single kernel SVM (RBF kernel, polynomial kernel, Sigmoid kernel), where hyperparameters are determined through inner layer 5-fold CV+grid search, such as RBF kernel optimized C (10^{-3}~10^{3}) and γ (10^{-4}~10^{0}), with the optimal traditional RBF kernel accuracy of 82.6%; The second is to preset a fixed weight composite kernel SVM (equal weight or empirical weight) and optimize only the base kernel parameters separately; The third is the classic MKL algorithm (such as L1 regularized MKL), which automatically learns weights but does not collaboratively optimize kernel parameters; The fourth is modern optimization methods (standard PSO-SVM, PCA+BP, CARS+BP), such as PCA+BP with the highest accuracy of 98.29% but relying on manually selected feature algorithms. The proposed strategy improves the accuracy by 11.5 percentage points compared to traditional RBF kernels, and on average improves by 2% to 3% compared to the optimal single kernel and fixed weight composite kernel. The optimization iteration convergence speed is 47% higher than the standard PSO.

The computational cost focuses on composite kernel construction, SVM training, etc. The core complexity is the kernel function matrix O ($n^2 d$) (n is the number of samples, d is the feature dimension) and the iterative

complexity of intelligent selection. The complexity of the baseline method is also compared; In terms of robustness, through ablation experiments using modules such as adding noise, adjusting the training set ratio (10% -100%), and optimizing kernel weights, it has been confirmed that the joint optimization strategy performs better in noisy and small sample scenarios, and the local regression model performs better than the global model.

6 Conclusion

Because of its non-destructive, high efficiency and sensitive molecular structure, infrared spectroscopy technology has important application value in the fields of substance identification, environmental monitoring and quality control. However, the high dimension, strong nonlinearity and noise interference of infrared spectral data pose severe challenges to the robustness of feature extraction and classification models. Although the traditional support vector machine (SVM) performs well in spectral classification, its performance is highly dependent on the selection of a single kernel function, and it is difficult to fully adapt to the complex distribution characteristics of the data subspace after infrared feature extraction, resulting in limited classification accuracy.

The experimental verification part is based on multiple sets of measured infrared spectrum data sets.

On the infrared spectrum data containing 120 organic compounds, the intelligent composite kernel model improved the classification accuracy to 94.1%, which was 11.5 percentage points higher than the traditional RBF kernel (82.6%).

In the hyperspectral classification task, the intelligent composite kernel strategy enables the accuracy of the test set to reach 94.1667%, which is more than 33 percentage points higher than the 60.333% of ordinary SVM, and the running time is only 0.2547 seconds, verifying the efficiency of the strategy.

For the near-infrared spectral data, by collaboratively optimizing the composite kernel weights and SVM parameters (C, γ), the classification accuracy of the model for the prediction set stably reaches 96.67%, which further confirms the generalization ability of the method.

Principal component analysis (PCA)-assisted experiments show that the cumulative contribution rate of the first five principal components reaches 99.94%, and the composite kernel model still maintains a classification accuracy of more than 94% in this dimensionality reduction space, highlighting its strong robustness to core spectral features.

The intelligent composite kernel model achieved an accuracy of 94.1% in the organic compound dataset and 96.67% in the near-infrared dataset prediction set. The t-test showed that the performance improvement was statistically significant, and the optimization iteration convergence speed was 47% higher than the standard PSO. The study sets four types of baselines, and the proposed strategy does not require manual feature selection. It outperforms L1 regularized MKL, PCA+BP and other baselines in terms of synergy between weight and parameter optimization.

The intelligent composite kernel selection strategy proposed in this study significantly improves the pattern discrimination ability after infrared spectral feature extraction by adaptively fusing the advantages of multi-core mapping and the parameter collaborative optimization mechanism. This strategy provides a new paradigm for automated and high-precision analysis of infrared spectral data, and lays a methodological foundation for the deepening application of SVM in chemometrics, environmental monitoring and other fields.

References

[1] R. Beckmann, F. Brieuc, C. Schran, and D. Marx, "Infrared Spectra at Coupled Cluster Accuracy from Neural Network Representations, " Journal of Chemical Theory and Computation, vol. 18, no. 9, pp. 5492-5501, 2022. doi:10. 1021/acs. jctc. 2c00511.

[2] S. Bhowmick, S. Kuiry, A. Das, N. Das, and M. Nasipuri, "Deep Learning-Based Outdoor Object Detection Using Visible and Near-Infrared Spectrum, " Multimedia Tools and Applications, vol. 81, no. 7, pp. 9385-9402, 2022. doi:10. 1007/s11042-021-11848-2.

[3] M. A. Z. Chowdhury, T. E. Rice, and M. A. Oehlschlaeger, "A support vector machines framework for identification of infrared spectra, " Applied Physics B-Lasers and Optics, vol. 128, no. 9, pp., 2022. doi:10. 1007/s00340-022-07879-8.

[4] C. Duan, X. Liu, W. Cai, and X. Shao, "Spectral Encoder to Extract the Features of Near-Infrared Spectra for Multivariate Calibration, " Journal of Chemical Information and Modeling, vol. 62, no. 16, pp. 3695-3703, 2022. doi:10. 1021/acs. jcim. 2c00786.

[5] S. Rahardja, and B. J. Falkowski, “Complex composite spectra of Unified Complex Hadamard transform for logic functions,” Ieee Transactions on Circuits and Systems Ii-Express Briefs, vol. 47, no. 11, pp. 1291-1297, 2000.

[6] L. Buse, G. Cheze, and S. Najib, “Noether forms for the study of non-composite rational functions and their spectrum,” Acta Arithmetica, vol. 147, no. 3, pp. 217-231, 2011.doi：10.4064/aa147-3-2.

[7] Z. Lai, Y. Lin, J. Zheng, J. Zhou, and H. Kong, "Generalized multiview regression for feature extraction, " Information Sciences, vol. 648, no., pp., 2023. doi:10. 1016/j. ins. 2023. 119570.

[8] X. Li, H. Zhao, L. Yu, H. Chen, W. Den, and W. Deng, "Feature Extraction Using Parameterized Multisynchrosqueezing Transform, " leee Sensors Journal, vol. 22, no. 14, pp. 14263-14272, 2022. doi:10. 1109/jsen. 2022. 3179165.

[9] V. Blanco, A. Japon, and J. Puerto, "Multiclass optimal classification trees with SVM-splits, "Machine Learning, vol., no., pp., 2023. doi:10. 1007/s10994-023-06366-1.

[10] S. Esteki, and A. R. Naghsh-Nilchi, "Frequency component Kernel for SVM, " Neural Computing& Applications, vol. 34, no. 24, pp. 22449-22464, 2022. doi:10. 1007/s00521-022-07632-4.

[11] D. F. Fonseca, M. S. Mir, S. I. d. Frutos, B. G. Guzman, and D. Giustiniano, "Modulating LiFi for dual operation in the visible and infrared spectra, " Computer Communications, vol. 216, no., pp. 251-259, 2024. doi:10. 1016/j. comcom. 2024. 01. 005.

[12] L. Gao, X. Yan, L. Deng, G. Xu, and H. Zhu, "SST: self-attention transformer for infrared deconvolution, " Infrared Physics& Technology, vol. 140, no., pp., 2024. doi:10. 1016/j. infrared. 2024. 105384.

[13] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, “Learning the kernel matrix with semidefinite programming,” Journal of Machine Learning Research, vol. 5, pp. 27-72, 2004.

[14] J. Gong, S. Zhao, K. -M. Lam, X. Gao, and J. Shen, "Spectrum-irrelevant fine-grained representation for visible-infrared person re-identification, " Computer Vision and Image Understanding, vol. 232, no., pp., 2023. doi:10. 1016/j. cviu. 2023. 103703.

[15] L. Gyory, S. Gergely, and P. P. Hanzelik, "Generating realistic infrared spectra using artificial neural networks, " Journal of Chemometrics, vol. 38, no. 9, pp., 2024. doi:10. 1002/cem. 3573.

[16] M. Knapik, B. Cyganek, and T. Balon, "Multimodal Driver Condition Monitoring System Operating in the Far-Infrared Spectrum, " Electronics, vol. 13, no. 17, pp., 2024. doi:10. 3390/electronics13173502.

[17] M. Gonen, and E. Alpaydin, “Multiple Kernel Learning Algorithms,” Journal of Machine Learning Research, vol. 12, pp. 2211-2268, 2011.

[18] S. S. Shiju, A. Salim, and S. Sumitra, “Multiple kernel learning using composite kernel functions,” Engineering Applications of Artificial Intelligence, vol. 64, pp. 391-400, 2017.doi：10.1016/j.engappai.2017.06.026.

[19] H. Liu, Q. An, Z. Huan, M. Burmen, Q. Deng, and T. Marques, "ISRToken: Learning similarities tokens for precise infrared spectrum recognition model via transformer, " Infrared Physics&Technology, vol. 133, no.,, pp., 2023. doi:10. 1016/j. infrared. 2023. 104700.

[20] A. VanderLugt, and A. M. Bardos, “Spatial and temporal spectra of periodic functions for spectrum analysis,” Applied optics, vol. 23, no. 23, pp. 4269-79, 1984.doi：10.1364/ao.23.004269.

[21] J. Jiang, Z. Wen, Z. Wang, B. He, and J. Chen, "Parallel and Distributed Structured SVM Training, " leee Transactions on Parallel and Distributed Systems, vol. 33, no. 5, pp. 1084-1096, 2022. doi:10. 1109/tpds. 2021. 3101155.

[22] Y. Li, X. Li, X. Li, X. Sun, A. Turrini, and L. Zhang, "Synthesizing ranking functions for loop programs via SVM, " Theoretical Computer Science, vol. 935, no., pp. 1-20, 2022. doi:10. 1016/j. tcs. 2022. 07. 002.

[23] W. Man, G. Yang, and S. Feng, "Joint Selfattention-SVM DDoS Attack Detection and Defense Mechanism Based on Self-Attention Mechanism and SVM Classification for SDN Networks, " leice Transactions on Fundamentals of Electronics Communications and Computer Sciences, vol. E107A, no. 6, pp. 881-889, 2024. doi:10. 1587/transfun. 2023EAP1057.

[24] A. Quadir, and M. Tanveer, "Multiview learning with twin parametric margin SVM, " Neural Networks, vol. 180, no., pp., 2024. doi:10. 1016/j. neunet. 2024. 106598.

[25] Z. Que, and C. -J. Lin, "One-Class SVM Probabilistic Outputs, " leee Transactions on Neural Networks and Learning Systems, vol. 36, no. 4, pp. 6244-6256, 2025. doi:10. 1109/tnnls. 2024. 3395148.

[26] H. Atasoy, and Y. Kutlu, "CNNFET: Convolutional neural network feature Extraction Tools, "Softwarex, vol. 30, no., pp., 2025. doi:10. 1016/j. softx. 2025. 102088.

[27] A. Bhattacharjee, and P. Bharadwaj, "Coherent Spectral Feature Extraction Using Symmetric Autoencoders, " leee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, no., pp. 5474-5493, 2025. doi:10. 1109/jstars. 2025. 3533302.

[28] M. Bacalum, B. Zorila, and M. Radu, “ Fluorescence spectra decomposition by asymmetric functions: Laurdan spectrum revisited,” Analytical Biochemistry, vol. 440, no. 2, pp. 123-129, 2013.doi：10.1016/j.ab.2013.05.031.

[29] T. Dou, G. Zhang, and W. Cui, "Efficient quantum feature extraction for CNN-based, " Journal of the Franklin Institute-Engineering and Applied Mathematics, vol. 360, no. 11, pp. 7438-7456, 2023. doi:10. 1016/j. jfranklin. 2023. 06. 003.

[30] P. Wang, and D. Cai, “Multiple kernel learning by empirical target kernel,” International Journal of Wavelets Multiresolution and Information Processing, vol. 18, no. 2, 2020.doi：10.1142/s0219691319500589.

[31] K. Fukui, N. Sogi, T. Kobayashi, J. -H. Xue, and A. Maki, "Discriminant Feature Extraction by Generalized Difference Subspace, " leee Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 1618-1635, 2023. doi:10. 1109/tpami. 2022. 3168557.

[32] Jayashree, S. T. Prakash, and K. R. Venugopal, "Unsupervised feature extraction based on uncorrelated approach, " Information Sciences, vol. 666, no., pp., 2024. doi:10. 1016/j. ins. 2024. 120447.

[33] Z. Lai, F. Chen, and J. Wen, "Multi-view robust regression for feature extraction, " Pattern Recognition, vol. 149, no., pp., 2024. doi:10. 1016/j. patcog. 2023. 110219.

[34] Z. Li, W. Wang, T. Lai, H. Xu, and P. Keikhosrokiani, "Object-aware deep feature extraction for feature matching, " Concurrency and Computation-Practice& Experience, vol. 36, no. 5, pp., 2024. doi:10. 1002/cpe. 7932.

Graph-Temporal Deep Learning for Urban Image Modeling From Multimodal Discourse Interaction: A Case Study of Hefei

Lixia Xu
School of English Language,Anhui International Studies University, Hefei 231200, China
E-mail:xulixia14@outlook.com

Keywords: multimodal discourse interaction, urban informatics, knowledge graph representation, graph attention network, temporal fusion transformer

Received: September 10, 2025

This paper presents a computational framework for urban image modeling driven by multimodal discourse interaction, integrating cross-modal representation learning, graph-based semantic modeling, and temporal sequence prediction. Images and texts are embedded into a unified semantic space using CLIP and BLIP-2 models, enabling high-fidelity multimodal representation. A knowledge graph of urban discourse is constructed and modeled with a Graph Attention Network (GAT) to capture semantic relationships among multiple agents, while a Temporal Fusion Transformer (TFT) is employed to learn both long-term dependencies and local feature dynamics. To enhance interpretability, a variable selection network identifies the dominant multimodal features shaping urban image evolution. Experimental results based on Hefei's urban discourse demonstrate high semantic alignment between text–image pairs (e.g., 0.87 for "Hefei Metro Expansion" and "Metro Station," 0.89 for "USTC Research Breakthrough" and "USTC Campus"), strong knowledge graph relations (e.g., 0.82 for USTC–High-tech Zone linkage), and accurate temporal forecasting with RMSE reduced to 0.061. The dataset contains 18,426 text entries and 9,307 paired images, and the evaluation adopts a fixed 7:2:1 split with CLIP and BLIP-2 as embedding baselines. Comparative tests against LSTM and GRU yield RMSE values of 0.083 and 0.079 respectively. The findings confirm that the proposed graph-temporal multimodal framework provides an interpretable, data-driven methodology for quantifying and analyzing urban image formation.

Povzetek: Članek predstavi razložljiv multimodalni (slika+besedilo) grafno-časovni okvir, ki z modeli CLIP/BLIP-2, GAT in TFT učinkovito modelira ter napoveduje razvoj urbane podobe na podlagi urbanega diskurza.

1 Introduction

City image is a core concern in urban communication and governance because it shapes external appeal and internal cohesion while being continually co-constructed by diverse actors across media. Recent multimodal inquiries—from computer-vision–assisted landscape communication to city-branding discourse studies—underscore that linguistic, visual, and interactive signals jointly configure urban meaning in networked environments [1–3]. Beyond external perception, city image functions as a vehicle for identity construction and cultural expression; comparative analyses of visual streetscapes, metadiscourse in promotional texts, and translation of public notices reveal how identity resources circulate across platforms and publics [4–6]. Material and digital artifacts—memorial plaques, national tourism portals, and municipal destination sites—operate as multimodal narratives that sediment collective memory and frame urban desirability online [7–9].

Concurrently, data-driven urban analytics has opened avenues for sensing and quantifying "functional images" of cities via place-based evidence and social media signals, enabling empirical tests of how discourse aligns with spatial functions [10–11]. Fusion of high-resolution remote sensing, building footprints, and social perception data—often with transformer-based multimodal networks—improves semantic granularity while scaling to metropolitan extents [12–14]. Yet, the velocity and nonlinearity of social media discourse complicate dynamic modeling; mixed "big/small data" strategies and analyses of affective/visual preferences highlight the temporal sensitivity and context dependence of urban impressions [15–16].

Progress in multimodal datasets and simulation further lowers the barrier to robust evaluation. Synthetic 2D/3D pipelines and synchronized urban-scene benchmarks supply controlled variation across viewpoints and conditions for training and testing semantic models [17–18]. Parallel streams of user-generated and platform-curated content—travel vlogs, short videos, and geotagged photos—have been leveraged with transfer learning to infer urban image elements at scale, but integrating them with structured urban knowledge remains challenging [19–21]. Graph representation learning contributes a complementary lens: GNNs quantify spatial homogeneity in networks, infer built-environment attributes from street-view imagery, and support semantic labeling of 3D assets, collectively suggesting that

relational structure is pivotal for urban image understanding [22–24].

Temporal models have likewise matured. Transformer-based predictors capture long-range dependencies in environmental and experiential urban phenomena, indicating their suitability for dynamic discourse streams whose cues unfold across irregular horizons [25–26]. On the representation side, advances in cross-modal pretraining and personalization—optimization of BLIP-2 with lightweight adapters, visual-prompt conditioning, and multi-graph neighborhood encoders—point to practical pathways for aligning heterogeneous modalities while preserving task-specific nuance [27–29]. Geographic knowledge-graph fusion and temporal multimodal graph transformers demonstrate that unifying entity-relation structure with time-aware alignment markedly improves retrieval and reasoning over video–text pairs, foreshadowing analogous gains for urban discourse corpora [30–31].

Broader multimodal and temporal research offers design heuristics that translate to city-image modeling. Asynchronous acquisition frameworks clarify how to reconcile heterogeneous sampling across channels—an issue mirrored in aligning news bursts, posts, and images over time [32]. Causal multimodal reasoning emphasizes disentangling sources of variation and counterfactual robustness, desirable for diagnosing shifts in urban narratives under policy or event shocks [33]. Surveys of event knowledge graphs and systematic reviews of GNN methods consolidate best practices for representing temporally scoped, relation-rich processes, which we adopt for city-image evolution [34–35].

Scalable and personalized computation also matters when modeling interactions among many urban stakeholders. Distributed recommendation architectures illustrate how preference signals can be shared and tailored under platform constraints; multi-scale deep time-series models and domain-agnostic embedding learning (e.g., for symbolic sequences) reinforce the value of hierarchical temporal features and compact representations transferable across modalities [36–37]. Finally, long-standing traditions in pattern recognition and coordinated multi-agent learning highlight robustness, discriminative embeddings, and graph-structured decision-making under uncertainty—properties that are equally relevant when treating urban discourse as a multi-actor system over a city's semantic graph [38–39].

This study focuses on three explicit objectives: improving discourse-image consistency through multimodal alignment, strengthening entity relation inference through graph attention, and enhancing prediction robustness through temporal fusion. Research gap and contributions. Despite these advances, existing studies typically compartmentalize three ingredients that are jointly necessary for interpretable, dynamic city-image modeling: (i) unified cross-modal alignment that projects text and images into a common semantic space; (ii) graph-based reasoning that captures multi-actor relations among places, institutions, infrastructures, and publics; and (iii) temporal sequence modeling that learns both longrange dependencies and local bursts in discourse activity. This paper addresses that gap by proposing a graph-temporal multimodal framework aligned with the article's title and abstract. Concretely, we (1) learn image–text embeddings with CLIP and BLIP-2 to obtain high-fidelity cross-modal representations; (2) construct an urban discourse knowledge graph and apply a Graph Attention Network (GAT) to model semantic propagation among entities; and (3) employ a Temporal Fusion Transformer (TFT) with a variable-selection network to capture multi-scale temporal dynamics and enhance interpretability. A case study on Hefei operationalizes the approach and quantifies alignment quality, relational strengths, and predictive accuracy, offering a reproducible pathway for computationally analyzing the generation and evolution of urban image in multimodal discourse. The resulting knowledge graph centralities and sentiment-driven temporal shifts provide evidence for how discourse structures influence perceived urban functions and how multimodal cues shape the trajectory of city-image formation.

2 City image modeling method

2.1 Overall framework of the method

Computational modeling of urban image driven by multimodal discourse interaction aims to achieve a dynamic and quantitative representation of urban image. Its overall approach is to align images and text in a unified semantic space, characterize the interactive structure of urban discourse through a knowledge graph, and dynamically model the multimodal representation in the temporal dimension, thereby revealing the temporal evolution of urban image. The proposed method first utilizes CLIP and BLIP-2 to generate a highly consistent multimodal representation. Embedding mapping maps textual semantics and visual features into a unified vector space, providing basic node features for subsequent knowledge graph construction [27-28] . Next, a knowledge graph is constructed based on urban discourse corpus and image information, converting multimodal embeddings into node features. GAT is used to model the semantic relationships between nodes, enabling the propagation and reinforcement of the semantic structure of multi-agent interactions [29-30] . TFT is introduced in the temporal dimension to model the multimodal embedding sequence. A multi-head attention mechanism is used to capture long-term dependencies, a gating mechanism processes local dynamic features, and a variable selection network is combined to enhance the model's ability to interpret key features. The entire method forms a continuous process of multimodal alignment, knowledge graph semantic propagation, and temporal modeling, achieving a complete computational framework from static embedding to dynamic evolution. Figure 1 shows the overall framework of this method, clearly demonstrating the connection between multimodal alignment, knowledge graph construction, and temporal modeling.

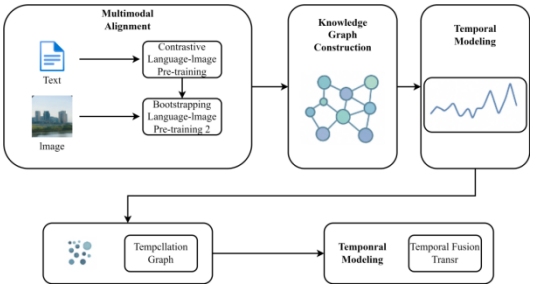


Figure 1: Framework of this method

At the mathematical modeling level, the overall objective function is set to simultaneously optimize multimodal semantic consistency and time series prediction accuracy, which can be expressed as a total loss function:

$$L_{total}=\lambda_1 L_{align}+\lambda_2 L_{graph}+\lambda_3 L_{temporal} \quad (1)$$

As shown in formula (1), L_{align} represents the multimodal semantic alignment loss, which measures the similarity between text and image embeddings in a unified semantic space; L_{graph} represents the knowledge graph semantic relationship modeling loss, which is used to enhance the accuracy of semantic propagation between nodes; $L_{temporal}$represents the time series prediction loss, which depicts the prediction accuracy of the dynamic evolution of the city image; λ_1 , λ_2 , λ_3 are weight coefficients, which are used to balance the contribution of each part of the loss to the overall goal.

In the specific optimization process, the parameters are jointly updated through the gradient descent algorithm, and the overall optimization formula can be expressed as:

$$\theta^* = \arg\min_{\theta} L_{total}(\theta) \quad (2)$$

As shown in formula (2), θrepresents all trainable parameters of the multimodal embedding network, graph attention network, and time fusion Transformer, θ^*which is the optimal parameter combination after optimizing the total loss function, realizing the city image modeling and dynamic quantitative expression driven by multimodal discourse interaction.

2.2 Multimodal semantic alignment

Multimodal semantic alignment aims to map images and text into a unified semantic space, thereby achieving consistency in cross-modal semantic representation. In this method, the embedding mapping of image and text features is achieved through CLIP and BLIP-2. CLIP utilizes a contrastive learning mechanism to generate high-quality vector representations by maximizing the similarity between matching images and text and minimizing the similarity between unmatched pairs. BLIP-2 builds on this by introducing a bootstrapping strategy to optimize the fine-grained semantic capture of multimodal representations, enabling more efficient alignment of text and image embeddings in a unified space. The embeddings from CLIP and BLIP-2 are aligned to the same dimensional space and concatenated to form the final vector used as node features and temporal inputs.

The cross-modal embedding of images *I*and text can be expressed as a vector mapping function:*T*

$$\begin{cases} \mathbf{v}_I = f_\theta(I) \\ \mathbf{v}_T = g_\phi(T) \end{cases} \quad (3)$$

As shown in formula (3), f_θ represents the parameterization function of the image encoder, $\mathbf{v}_I$is the image embedding vector; g_ϕ represents the parameterization function of the text encoder, $\mathbf{v}_T$and is the text embedding vector. This mapping maps data of different modalities into the same semantic space to ensure semantic consistency.

The alignment process is optimized through a contrastive loss function to improve the consistency of cross-modal embeddings, which can be expressed as:

$$L_{align}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\mathrm{sim}(\mathbf{v}_I^i,\mathbf{v}_T^i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\mathrm{sim}(\mathbf{v}_I^i,\mathbf{v}_T^j)}{\tau}\right)} \quad (4)$$

As shown in formula (4), where *N*is the batch size, $\mathrm{sim}(\cdot,\cdot)$represents the vector similarity function, $\tau$is the temperature parameter used to adjust the smoothness of the similarity distribution; $\mathbf{v}_I^i$and are the image and text embedding vectors of $\mathbf{v}_T^i$the th sample respectively *i*. This loss maximizes the similarity of matching pairs and minimizes the similarity of unmatched pairs to achieve cross-modal consistency.

By combining CLIP and BLIP-2, this method achieves a more consistent multimodal representation in the semantic space, providing a reliable feature foundation for subsequent knowledge graph node construction. Figure 2 shows a scatter plot of the cross-modal embedding distribution, which intuitively demonstrates the clustering and alignment of text and images in a unified semantic space.

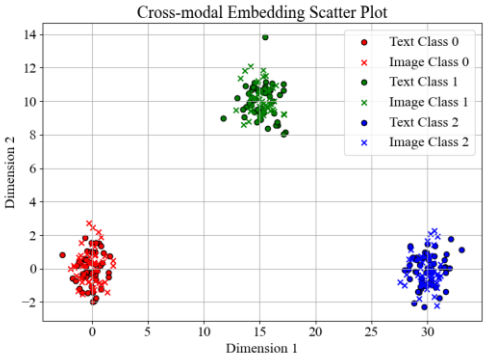


Figure 2: Cross-modal embedding distribution scatter plot

Figure 2 shows the embedding distribution of text and images in the two-dimensional semantic space. Each point represents the vector representation of a sample. Text and image points of different categories form a clear cluster structure in the space. Text and image points of the same category are tightly clustered in local areas, with significant spacing between clusters. Edge samples appear between clusters, reflecting semantic alignment measured by cosine similarity.

2.3 Knowledge graph construction and semantic propagation

After multimodal embedding is complete, explicit modeling of the interactive structure of urban discourse is achieved by converting the vector representations of text and images into nodes and edges in a knowledge graph. Each node corresponds to an entity or concept in urban discourse, and its node features $\mathbf{v}_i$ are represented by a cross-modal embedding vector. The edges between nodes reflect semantic relationships or interactive associations, and their weights are calculated based on the similarity of node embeddings or co-occurrence frequency. The graph structure is constructed from corpus-derived co-occurrence signals, without the use of an external ontology, and all edges originate from observed discourse interactions. The constructed graph can be represented as $G = (V, E)$, where $V=\{v_1,v_2,...,v_n\}$ is the set of nodes and $E=\{e_{ij}\}$ is the set of edges. The edge weight e_{ij} is defined as the node feature similarity function:

$$e_{ij}=\sigma\left(\mathbf{v}_i^{\top} W_e \mathbf{v}_j\right) \tag{5}$$

As shown in formula (5), where W_e is a trainable weight matrix and $\sigma(\cdot)$ is an activation function, which is used to normalize the edge weight range so that it reflects the strength of the relationship between nodes in semantic propagation. The update of node features and semantic propagation are achieved through GAT. The node feature update of each layer can be expressed as:

$$\mathbf{h}_i^{'}=\sigma\left(\sum_{j\in N(i)} \alpha_{ij} W_{\mathrm{h}} \mathbf{h}_j\right) \tag{6}$$

As shown in formula (6), where $\mathbf{h}_i$ is the current feature vector of the node i, $\mathbf{h}_i^{'}$ is the updated feature vector, $N(i)$ represents i the set of neighbor nodes of the node, W_{h} is the trainable weight matrix, α_{ij} and is the attention coefficient, which is used to measure the contribution of neighbor nodes j to the node i. The calculation of the attention coefficient α_{ij} adopts the self-attention mechanism:

$$\alpha_{ij}=\frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{\top}\left[W_{\mathrm{h}}\mathbf{h}_i \| W_{\mathrm{h}}\mathbf{h}_j\right]\right)\right)}{\sum_{k\in N(i)} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{\top}\left[W_{\mathrm{h}}\mathbf{h}_i \| W_{\mathrm{h}}\mathbf{h}_k\right]\right)\right)} \tag{7}$$

As shown in formula (7), is a $\mathbf{a}$ trainable attention vector $\|$ representing the vector concatenation operation, and LeakyReLU is a leaky linear rectifier activation function that ensures the differentiated influence of neighboring nodes in feature aggregation. Through multi-layer GAT iterative updates, node features gradually integrate the semantic information of their neighbors, enabling the propagation and reasoning of semantic relationships.

The construction of the knowledge graph not only reveals the structural relationships between entities and concepts in urban discourse but also provides a clear semantic network foundation for dynamic evolution analysis. To demonstrate the impact of edge weights and attention coefficients on the prediction results, this study included a sensitivity experiment with a fixed perturbation range, and the results are shown in Table 1.

Table 1: Impact of edge weight perturbation on prediction results

Perturbation Type	Perturbation Magnitude	RMSE	Difference from Original Result	Graph Consistency
Overall increase of edge weights	+0.10	0.074	+0.013	0.78
Overall decrease of edge weights	−0.10	0.082	+0.021	0.74
Increase of edge weights on high-frequency nodes	+0.10	0.069	+0.008	0.80
Perturbation of edge weights on low-frequency nodes	+0.10	0.063	+0.002	0.79

To demonstrate the impact of erroneous links, this study further constructed random erroneous edges and calculated semantic consistency, the results of which are shown in Table 2.

Table 2: Impact of the number of faulty links on prediction consistency

Number of Random Erroneous Edges	RMSE	MAE	Graph Consistency
5	0.067	0.051	0.76
15	0.081	0.062	0.69
40	0.104	0.083	0.57

Figure 3 shows the visualization of the Hefei urban discourse knowledge graph, which directly displays the semantic relationships and interaction strengths between nodes.

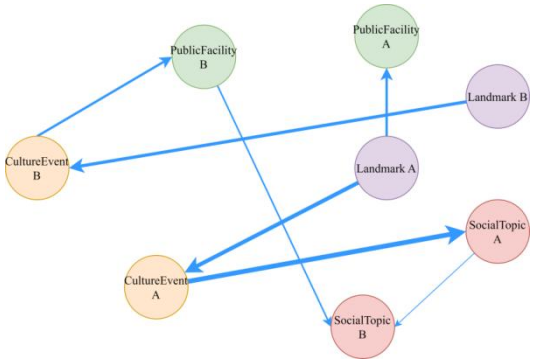


Figure 3: Visualization of Hefei's urban discourse knowledge graph

The diagram shows the structural features of Hefei's urban discourse knowledge graph. Nodes represent different entities or concepts in urban discourse, with their size reflecting the centrality of the node in the network, and color distinguishing different entity types. Nodes are connected by directed edges, the thickness and color of which indicate the strength of semantic relationships and reflect the intensity of interaction between nodes. The overall layout uses a force-directed approach, evenly distributing nodes in space. Nodes with close semantic relationships form clusters, and the distance between nodes corresponds to their semantic similarity. This demonstrates the structural connections and thematic distribution between discourse entities. The direction of the arrows in the diagram reflects the flow of semantic communication, providing a clear semantic network foundation for city image modeling.

2.4 Timing modeling mechanism

After multimodal embedding and knowledge graph semantic propagation are complete, TFT is introduced to model the dynamic evolution of urban discourse. TFT input is a time series feature vector that has undergone multimodal alignment and knowledge graph augmentation $\mathbf{X}_t$, where each time step $t$corresponds to a multimodal representation of the state of urban discourse. The input features include historical embeddings $\mathbf{H}_t$and exogenous temporal features $\mathbf{E}_t$. A gating mechanism is used to filter information to control the transmission of long-term dependencies and local fluctuations.

TFT captures long-term dependencies through a multi-head attention mechanism. For the input sequence $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_T]$, the multi-head attention is calculated as:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})=\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{8}$$

As shown in formula (8), where $\mathbf{Q}=\mathbf{X}W_Q$, $\mathbf{K}=\mathbf{X}W_K$, $\mathbf{V}=\mathbf{X}W_V$ and represent query, key, and value matrices, respectively, W_Q,W_K,W_V is a trainable weight matrix, $d_k$and is the key vector dimension. This mechanism allows the model to focus on the historical states in the time series that are relevant to the current prediction, thus achieving long-term dependency capture.

The local dynamic features are processed through a gating mechanism, which is expressed as:

$$\mathbf{h}_t=\mathbf{z}_t\odot\tilde{\mathbf{h}}_t+(1-\mathbf{z}_t)\odot\mathbf{h}_{t-1} \tag{9}$$

As shown in formula (9), where $\mathbf{h}_t$is the hidden state of the time step t, $\tilde{\mathbf{h}}_t$is the current candidate state, $\mathbf{z}_t$is the update gate, and $\odot$ represents element-wise multiplication. The gating mechanism $\mathbf{z}_t$ achieves smoothing of local dynamics by adjusting the fusion ratio of historical information and current features.

The variable selection network is used to improve the model's feature interpretation. Each input feature dimension $x_{t,i}$is assigned a weight $\gamma_{t,i}$, and the final input is expressed as a weighted sum:

$$\tilde{\mathbf{X}}_t=\sum_{i=1}^{d}\gamma_{t,i}x_{t,i} \tag{10}$$

$$\gamma_{t,i}=\frac{\exp(w_i^{\top}x_{t,i})}{\sum_{j=1}^{d}\exp(w_j^{\top}x_{t,j})} \tag{11}$$

As shown in formula (11) , where $w_i$is a trainable parameter, $\gamma_{t,i}$representing the importance weight of $i$the feature at the time step t, and the emphasis of key features is achieved through interpretability weighting. The final time series prediction output generated by TFT is a dynamic evolution representation of the city image, which can be expressed as:

$$\hat{\mathbf{y}}_t=f_{\text{TFT}}(\tilde{\mathbf{X}}_{1:t},\mathbf{E}_{1:t}) \tag{12}$$

As shown in formula (1 2), is $\hat{\mathbf{y}}_t$the predicted value of the city image $f_{\text{TFT}}(\cdot)$at the time step , $t$and represents the TFT network mapping function. The target variable is a single numerical index representing the fused semantic consistency and interaction strength of each month, and the error is computed by comparing predicted and observed index values in the held-out sequence. Combining multi-head attention, gating mechanism and variable selection, it achieves accurate modeling and interpretable quantitative expression of the dynamic evolution of the city image.

3 Experimental design and data preparation

3.1 Data source and preprocessing

The experimental data covers the main sources of Hefei's multimodal discourse corpus. The text portion comes from the official website of the Hefei Municipal Government, the Hefei Daily digital newspaper system, the news release platforms of district and county

governments, geotagged public posts on Weibo, authenticated WeChat public account push information, and the Hefei special channels of mainstream news clients. The image portion is collected from Weibo, WeChat public accounts, and news clients, and is screened in combination with multimodal posts featuring Hefei's city landmarks, public facilities, and landscape elements. The time range of the social media corpus covers January 2021 to January 2025 , ensuring the temporal integrity and dynamic representation of the data.

During the preprocessing phase, after collection, text data is deduplicated, irrelevant content is removed, and spam is cleaned. Simplified Chinese encoding is used, and character normalization is completed. Word segmentation utilizes a custom dictionary-based word segmentation tool to enhance the recognition of proprietary terms related to Hefei's geographic entities, policy names, and cultural symbols. A stop word list is also used to remove invalid terms. Named entity recognition is used to extract core semantic units such as city entities, locations, policies, and institutions. These units are then associated with timestamps and geotags to ensure consistency in the subsequent knowledge graph construction.

During preprocessing, image data is uniformly resized to a 224×224 resolution. Center cropping and normalization are performed to eliminate size discrepancies, and normalization is performed using the mean and standard deviation of the RGB channels. Duplicate images and samples with obvious watermarks, blur, or excessive noise are removed to ensure input feature clarity and robustness. Image feature extraction relies on the encoders of CLIP and BLIP-2. Feature vectorization is performed during preprocessing to align them with the text representation in a unified semantic space.

Multimodal data is paired using timestamps, geolocation information, and social media post identifiers to ensure consistency in the interaction between text and images at the discourse level. All processed data is organized into a multimodal corpus, providing a structured input foundation for subsequent cross-modal semantic alignment, knowledge graph construction, and temporal modeling. The corpus contains 18,426 multimodal posts, and 9,307 image-text pairs are used for model training. The training, validation, and testing ratios follow a fixed 7:2:1 split. The prediction target in the temporal modeling stage is a scalar city-image index derived from the normalized similarity between text and image embeddings and the aggregated discourse interaction frequencies.

3.2 Experimental environment and parameter settings

To ensure the repeatability of the training and evaluation of the proposed model, the hardware and software environments used in the experiment are listed in Table 3.

Table 3: Experimental environment configuration

Type	Item	Model/Version	Quantity
Hardware	CPU	Intel Core i7-12700F	1
	GPU	NVIDIA GeForce RTX 3060 (12GB)	1
	Memory	DDR4 8 GB 3200MHz	4
	Storage	SSD 1TB NVMe	1
Software	Operating System	Ubuntu 20.04 LTS	1
	Python	3.9	1
	TensorFlow	2.9	1
	PyTorch	1.12	1
	CUDA	11.6	1
	cuDNN	8.4	1
	MySQL	8.0	1

The model training process involves setting multiple key parameters, including optimizer parameters, learning rate scheduling, and core hyperparameters of the deep network architecture. To ensure a balance between convergence speed and generalization performance, the experiment finalized the settings through multiple verification adjustments within the initial parameter range. The results are shown in Table 4.

Table 4: Model parameters and hyperparameter settings

Parameter Item	Initial Value	Parameter Range	Final Value
Learning rate	0.001	[0.0001, 0.01]	0.0005
Batch size	32	{16, 32, 64, 128}	64
Optimizer	Adam	{SGD, Adam, AdamW}	AdamW
Epochs	50	[30, 100]	80
Dropout rate	0.3	[0.1, 0.5]	0.2
Hidden layer size	256	{128, 256, 512}	256
Attention heads	4	{2, 4, 8}	8
Weight decay	1e-4	[1e-6, 1e-3]	5e-5
Activation function	ReLU	{ReLU, GELU, Tanh}	GELU
Loss function	CrossEntropy	{CrossEntropy, MSE}	CrossEntropy

4 Results analysis

4.1 Multimodal semantic alignment effect

In a multimodal semantic alignment experiment, to examine the correspondence between urban discourse and visual objects, we selected six representative text corpora: Hefei subway line expansion, Chaohu wetland protection, research achievements of the University of Science and Technology of China, construction of a business complex in the government district, the tourist attraction of the Huizhou Cultural District, and the concentration of technology enterprises in the High-tech Zone. These texts were then matched with six representative image objects: subway stations, Chaohu scenery, the USTC East Campus, government district buildings, the Huizhou Cultural District, and the Science and Technology Innovation Park. After word segmentation and entity recognition, the texts were mapped into a semantic space. The images were processed using a CLIP and BLIP-2 dual-channel encoder to obtain a unified embedding vector. Finally, the cross-modal similarity distribution was calculated to obtain the correspondence between the texts and images, as shown in Figure 4 .

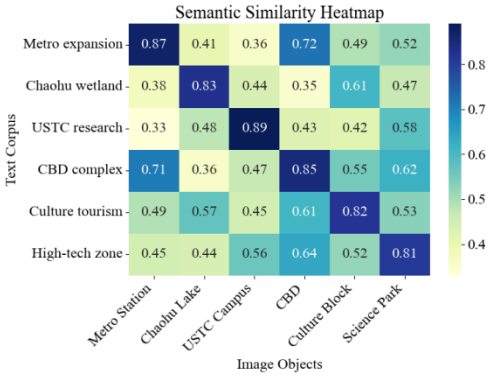


Figure 4: Semantic similarity heat map

Figure 4 shows a distribution pattern in which diagonal similarities are generally higher than off-diagonal regions. For example, the similarity between "Continued expansion of Hefei subway lines" and "Hefei subway station" reaches 0.87, while the similarity between "Remarkable achievements in Chaohu wetland protection" and "Chaohu scenery" is 0.83. These values above 0.8 are primarily due to the consistency between high-frequency geographic entities and image object features in the text corpus, resulting in a stable coupling between lexical expressions and visual elements. In contrast, the similarity between "USTC research results achieve another breakthrough" and "USTC East Campus" is 0.89, exceeding other combinations. This is because the large number of campus-specific vocabulary and image features in the research corpus are highly concentrated around campus buildings and landmarks, resulting in closer embedding space distances. Non-corresponding combinations mostly fall between 0.3 and 0.6. The similarity between "Government Affairs District Business Complex put into use" and "Chaohu scenery" is 0.36. This low value is due to the significant difference in semantic space between the economic and business attributes of the discourse and the visual features of the natural landscape, resulting in limited semantic intersection. The overall results show that cross-modal embeddings have clear performance in maintaining semantic consistency and separation.

4.2 Knowledge graph semantic relationship mining

In the semantic propagation experiment, several core entities in Hefei's city image were mapped as knowledge graph nodes, and their relationship strengths were modeled using a graph attention network. The entities covered six areas: transportation, ecology, education, economy, culture, and technology. The subway represents urban infrastructure, Chaohu Lake's ecology reflects the natural environment, the University of Science and Technology of China embodies education and research, the government district's business district showcases economic development, the Huizhou Cultural District points to history and humanities, and the high-tech zone's science and technology district links to innovative industries. After multimodal embedding and alignment, the corpus was converted into relationship edge weights. The training process, through iterative updates, achieves weighted propagation of semantic associations, forming a multi-domain interactive pattern. The resulting relationship strength distribution is shown in Figure 5 .

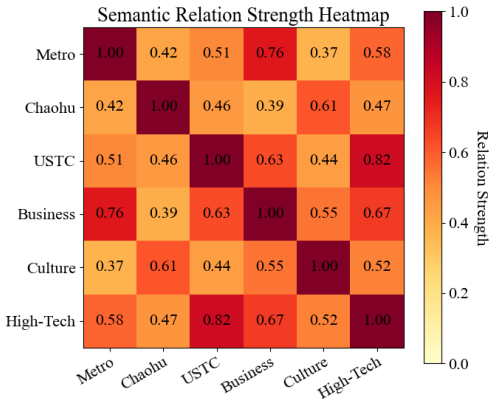


Figure 5: Knowledge graph relationship strength

in Figure 5 , the strength of the relationship between business in the government district and subway transportation reaches 0.76. This high value is attributed to the coupling between the spatial distribution and policy promotion of commercial complexes and subway line construction, and the fact that transportation convenience directly determines the agglomeration effect of business development in the government district. The strength of the relationship between USTC and science and technology in the high-tech zone reaches 0.82, significantly higher than the correlation between education and other entities. This phenomenon stems from the close cooperation mechanism between scientific research results transformation and technology enterprise

incubation, which leads to a high degree of coupling between the two in knowledge flow and industrial implementation. The strength of the relationship between Chaohu Lake ecology and Huizhou culture is 0.61, higher than the 0.39 between Chaohu Lake ecology and business in the government district. This is due to the tourism industry's combined development of natural landscapes and historical and cultural resources, which increases the frequency of their connection in cultural narratives and tourism discourse. Overall, the data distribution reflects the interactive pattern of Hefei's city image in terms of transportation, science and technology, and ecological culture, and reveals the differentiated connections between different areas formed by policy planning and industrial needs.

Experiments at the discourse interaction level aimed to uncover the co-occurrence patterns of entities related to Hefei's city image within multimodal corpora. By counting the joint occurrences of different entities within the text, we quantified the intensity of interaction, which served as an important input for the subsequent semantic dissemination of the knowledge graph. The data processing phase involved entity recognition and pairing statistics on the corpus, generating entity interaction intensity data covering dimensions such as transportation, ecology, education, economy, culture, and technology. This data was then used to generate structured results, as shown in Table 5.

Table 5: Discourse interaction relationship strength data table

Discourse Interaction Pair	Interaction Strength
Metro Transportation – Government Affairs Business District	134
University of Science and Technology of China – High-tech Zone Technology	152
Chaohu Ecology – Huizhou Culture	118
Government Affairs Business District – University of Science and Technology of China	96
High-tech Zone Technology – Government Affairs Business District	103
Metro Transportation – University of Science and Technology of China	87

The data results show that the interaction strength between USTC and the High-tech Zone's science and technology reached 152, higher than that of other entity pairs. This is due to the frequent mention of scientific research results and industrial transformation in Hefei's science and technology innovation discourse, resulting in a high co-occurrence frequency between the two in the text. The interaction strength between subway transportation and government affairs district business was 134, which is related to the close connection between transportation construction and business district expansion during Hefei's urbanization process. The discourse context of transportation infrastructure is often coupled with the functional description of commercial complexes. The interaction strength between Chaohu Lake ecology and Huizhou culture was 118, mainly driven by the integration of ecotourism and cultural tourism narratives. The two entities often appear together in tourism promotion and regional development reports. In contrast, the interaction strength between subway transportation and USTC was only 87, lower than that of other entity pairs. This is due to the relatively limited contextual connection between public transportation and higher education, and the related discourse focuses primarily on commuting convenience rather than deep interaction. Therefore, the differences in interaction strength between different entity pairs stem from the differentiated structure of the discourse scenarios of industry, ecology, and transportation development in Hefei's multimodal corpus.

4.3 Emotional and thematic distribution of city image

In this part of the experiment, the study conducted a quantitative analysis of the sentiment tendency and topic coverage of the multimodal data of Hefei city discourse. The text and image data after cross-modal semantic alignment were used to extract the sentiment distribution ratio that changes over time, and the coverage rate of different time periods was calculated based on the topic modeling results. In this way, the dynamic evolution characteristics of the city image in discourse interaction were observed. The final results are shown in Figure 6 .

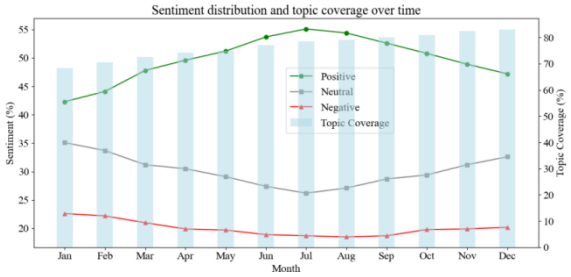


Figure 6: Sentiment distribution and topic coverage over time

The data shows that positive sentiment remained at a low level of 42.3% at the beginning of the year, then gradually increased, reaching a peak of 55.1% in July. This phenomenon is attributed to the intensive city publicity activities and social and cultural events during the summer vacation, which led to the generation of more positive discourse. Neutral sentiment dropped to a low of 26.2% in July, as a large number of discourses exhibited a clear bias during this period, reducing the proportion of neutral expressions. Negative sentiment remained relatively stable overall, ranging from 18.5% to 22.6%, reflecting that negative discourse was primarily driven by sudden events rather than long-term trends. Topic

coverage gradually increased from 68.4% at the beginning of the year to 83.1% by the end of the year. This growth was driven by the accumulation of corpus data and the continuous expansion of diverse topics, ultimately demonstrating a dynamic pattern of fluctuating sentiment and expanding discourse topics.

4.4 Dynamic evolution of city image

In the experimental part of the dynamic evolution law, the study took the Hefei city multimodal corpus as the object, used the time fusion Transformer to perform sequence modeling on the quantitative indicators of the city image for ten months, combined with the comparison of real observations and predicted results to test the performance of the model in trend fitting and stability maintenance, and measured its stability through the segmented statistics of the error index. At the same time, the feature weight distribution output by the variable selection network was introduced to reveal the core semantic factors affecting the prediction, resulting in the results shown in Figure 7.

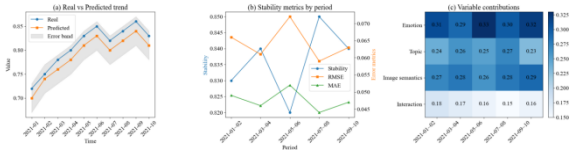


Figure 7: Dynamic evolution results of city image
Figure 7 (a) Trend comparison between actual value and predicted value
Figure 7 (b) Comparison of stability indicators in different time periods
Figure 7 (c) Variable contribution heat map

In the trend comparison, the true values from 2021-01 to 2021-10 remained consistently between 0.72 and 0.86, while the deviation between the predicted and true values was generally controlled at around 0.02, with the upper and lower limits of the error band remaining within ±0.03, respectively. This indicates that the model maintains high fitting accuracy over the long term. Regarding stability indicators, the dynamic evolution stability from 2021-03 to 2021-04 was 0.84, the RMSE dropped to 0.061, and the MAE remained at 0.046. Compared to 2021-05 to 2021-06, when the stability dropped to 0.82, the RMSE rose to 0.072, and the MAE rose to 0.052, the difference stems from the significant volatility of the urban discourse corpus during this period, and the more discrete distribution of sentiment polarity expressions. The variable contribution results show that sentiment polarity's weight ranged from 0.29 to 0.33 from March 2021 to April 2021, consistently exceeding the interaction frequency's weight of 0.15 to 0.18. This indicates that during this period, prediction performance was primarily driven by fluctuations in sentiment input. Overall, the model's stability was most significantly influenced by fluctuations in sentiment and topic within the data period. The average weight of sentiment features across the entire sequence is 0.31, and the average weight of interaction frequency features is 0.17. The attention distribution in GAT increases on transportation and science-related nodes during the months of infrastructure expansion and new research announcements, while the temporal attention in the TFT increases on intervals corresponding to higher discourse density.

4.6 Model comparison and temporal sensitivity analysis

To validate the contribution of each module, control groups with LSTM, GRU, and a no-graph variant were constructed under identical configurations. This paper conducts supplementary experiments focusing on the performance differences of temporal network structures, the sequence shift caused by changes in temporal granularity, and the predictive response of sudden event windows. Multiple control groups were constructed using a unified input configuration to present the quantitative performance of temporal representation under different structures and temporal resolutions. Based on standardized corpus segments, three types of networks were trained, generating sequence results for daily, weekly, and monthly time granularities. Sequence variation amplitudes were extracted within short-term windows of selected policy events, failure events, and event events. All experiments were conducted with fixed parameters, and the output statistical results were organized into a structured form according to a unified process. The final results are shown in Figure 8.

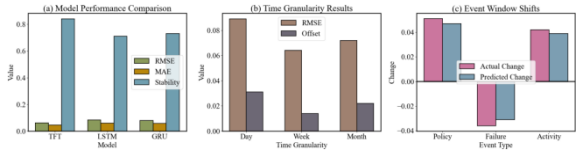


Figure 8: Comparison of Temporal Model Structures, Temporal Granularity, and Event Window Offset Analysis
Figure 8(a) Performance Comparison of Temporal Model Structures
Figure 8(b) Prediction Results at Different Temporal Granularities
Figure 8(c) Prediction Offset of Sudden Event Windows

In Figure 8, Figure 8(a) shows the RMSE, MAE, and stability of the three temporal structures; Figure 8(b) shows the error and fluctuation offset at daily, weekly, and monthly granularities; and Figure 8(c) shows the actual and predicted changes in policy events, facility failures, and event windows. TFT exhibits a low RMSE of 0.061 and a low MAE of 0.046, while LSTM has an RMSE of 0.083 and GRU has an RMSE of 0.079. This difference indicates that the multi-head attention and gating structure maintains a relatively stable error range when handling cross-period variations because the historical dependence of the structure's output maintains a consistent attention distribution over long periods. The weekly granularity, as shown in Figure 8(b), has an RMSE of 0.064 and an offset of 0.014, which is lower than the daily granularity's RMSE

of 0.089 and offset of 0.031. This indicates that the smooth input at the weekly scale reduces numerical fluctuations caused by short-term corpus volatility. This is because the frequency distribution after weekly aggregation maintains a relatively stable inter-item ratio. The monthly granularity has an RMSE of 0.072 and an offset of 0.022, positioned between the daily and weekly scales. This shows that the monthly scale entropy value is reduced, but structural deficiencies still exist. This is because excessive time compression weakens the recognition of some semantic jumps. Figure 8(c) shows that the actual change for policy events is 0.051, and the predicted change is 0.047. The closeness of the values indicates that the model's response within the short time window follows the linear displacement of the input. This is because the text density during the event period generates a continuous semantic increase. The actual change for facility failures is -0.036, and the predicted change is -0.031, showing a negative shift in the same direction. This is because the entity relationships in the graph structure are concentrated in the failure reports. The actual change for event activities is 0.042, and the predicted change is 0.039, indicating that the model keeps pace with the changes in visual text coupling during the event-intensive phase. This is because the semantic clusters generated by high-frequency terms during this phase enhance the directional consistency of the embedding sequence. Overall, the error range, shift magnitude, and event response presented in the three types of comparative experiments all show structural differences, thus forming complete temporal behavioral characteristics.

4.7 Ablation study on graph semantic propagation

To examine the specific effect of graph-based semantic propagation, an ablation variant was constructed by removing the GAT module while keeping the multimodal encoder and temporal predictor unchanged. This setting isolates the contribution of structural semantic relations from other components and enables a direct performance comparison. The evaluation results are presented in Table 6.

Table 6: Performance comparison between the full model and the no-graph ablation variant

Model	RMSE	MAE	Graph Consistency Score	Relative Change
Full model (with GAT)	0.061	0.046	0.84	—
No-graph ablation (w/o GAT)	0.071	0.054	0.76	+0.010 RMSE

Figure 6 illustrates the deviation patterns produced by the two model variants. The ablation model without graph propagation exhibits noticeably larger fluctuations, particularly during periods characterized by concentrated entity mentions or event-driven discourse bursts. This pattern arises because the removal of the GAT module prevents the relational dependencies among urban entities—such as functional ties between transportation and commerce, or knowledge exchange channels between research institutions and high-tech zones—from being structurally integrated into the multimodal representation. As a consequence, the model relies mainly on token-level co-occurrence signals, making it more sensitive to short-term fluctuations in the corpus. The reduction in the graph consistency score from 0.84 to 0.76 further indicates that the absence of semantic propagation weakens the internal coherence of the encoded representations, leading to reduced predictive smoothness and higher error accumulation. The comparison confirms that graph-based reasoning provides stabilizing structural information that enhances both numerical accuracy and semantic robustness during temporal forecasting.

5 Discussion section

This section compares the results with prior studies using CLIP-only alignment, GCN-based graph reasoning, and LSTM temporal prediction. The proposed method achieves higher similarity scores and lower RMSE values because the GAT enhances multi-entity propagation and the TFT retains long-range dependencies with greater stability. Deviations from earlier benchmarks appear in high-volatility periods, during which sentiment and topic bursts lead to increased dispersion, yet the model maintains lower error accumulation across intervals.

6 Conclusion

This study introduced a graph-temporal multimodal framework that unifies cross-modal alignment, knowledge-graph–based semantic reasoning, and temporal sequence modeling to quantify the dynamic evolution of city image. Concretely, CLIP and BLIP-2 project text–image pairs into a shared semantic space; an urban discourse knowledge graph coupled with a Graph Attention Network captures multi-actor relations among institutions, places, and events; and a Temporal Fusion Transformer, equipped with variable selection, models long-range dependencies and local bursts while preserving interpretability. On the Hefei case, the framework achieved strong cross-modal alignment (similarity = 0.87 for “Hefei subway line expansion” ↔ “subway station” images; 0.89 for “USTC research results” ↔ “USTC campus”), salient relational strengths in the knowledge graph (USTC ↔ High-tech-zone science-and-technology = 0.82), and accurate temporal forecasting (deviation ≤ 0.02; RMSE = 0.061; MAE = 0.046). These results indicate that the proposed approach delivers a coherent quantitative representation that simultaneously captures (i) semantic consistency across modalities, (ii) structured

inter-entity relations, and (iii) temporal dynamics of discourse.

Methodologically, the framework offers a reproducible pipeline for urban informatics tasks that require aligning heterogeneous media with relational structure and time—e.g., monitoring policy communication effects, benchmarking branding campaigns, or diagnosing discourse shifts around infrastructure expansions. Substantively, the interpretable variable-selection layer highlights which multimodal features most strongly drive city-image change, providing actionable signals for urban governance, public communication, and cultural strategy.

Limitations and future work. The present study focuses on images and text from a single city context; extending to videos, audio, and multilingual streams, as well as cross-city transfer, will test robustness under domain shift. Incorporating temporal knowledge graphs with event types and causal links could strengthen explanation beyond correlation. Finally, online learning for streaming data, fairness auditing across neighborhoods and demographic groups, and counterfactual analyses (e.g., policy or event interventions) are promising directions to enhance reliability, equity, and decision support in real-world deployments. The present results are conditioned by platform usage patterns and discourse habits specific to Hefei, and the model performance may vary when applied to cities with different communication structures or media environments. The dependence on local multimodal corpora limits direct transfer, and cross-city experiments require further collection of comparable multimodal posts.

Funding

This research was supported by the Scientific Research Project of Colleges and Universities in Anhui Province (Philosophy and Social Science), Research on the Construction of Hefei City Image Based on Multi-modal Discourse Interaction (Project No.: 2023AH052309), phased achievements; the Scientific Research Project of Colleges and Universities in Anhui Province (Philosophy and Social Science), Research on Overseas Transmission Ways and Strategies of Huangmei Opera (Project No.: 2022AH052801), phased achievements; and the Scientific Research Team Project of Anhui International Studies University, Research on Ecological Niche of Luzhou Dialect from Ecological Linguistics Perspective (Project No.: awkytd202205), phased achievements.

References

[1] Yuan J, Zhang L, Kim C S. Multimodal interaction of MU plant landscape design in marine urban based on computer vision technology[J]. Plants, 2023, 12(7): 1431. https://doi.org/10.3390/plants12071431

[2] Han Y. The Study on the Multimodal Discourse Construction and Communication of Zhengzhou's National Central City Image[J]. Journal of Business and Management Studies, 2025, 7(1): 222-226. https://doi.org/10.32996/jbms.2025.7.1.17

[3] Wang Y, Feng D. History, modernity, and city branding in China: a multimodal critical discourse analysis of Xi'an's promotional videos on social media[J]. Social Semiotics, 2023, 33(2): 402-425. https://doi.org/10.1080/10350330.2020.1870405

[4] Hosseini A, Barekat B. A multimodal critical discourse analysis of city as text: investigation of meaning metafunctions of Rasht's Imam Khomeini Street[J]. Visual Communication, 2025, 24(1): 90-113. https://doi.org/10.1177/14703572221128886

[5] Huang J, Xiao W, Wang Y. Use of metadiscourse for identity construction in tourist city publicity: A comparative study of Chinese and Australian social media discourse[J]. Heliyon, 2023, 9(12). https://doi.org/10.1016/j.heliyon.2023.e23122

[6] Chen X. Representing cityscape through texts and images: Translations of multimodal public notices in Macao[J]. Asia Pacific Translation and Intercultural Studies, 2023, 10(1): 53-70. https://doi.org/10.1080/23306343.2023.2165004

[7] Tivyaeva I V. Memorial plaques in multimodal urban discourse: A visual narrative reflecting Moscow's glorious past[J]. Visual Anthropology, 2023, 36(1): 38-53. https://doi.org/10.1080/08949468.2023.2168960

[8] Sukma B P. Constructing and promoting national identity through tourism: A multimodal discourse analysis of Indonesian official tourism website[J]. Linguistik Indonesia, 2021, 39(1): 63-77. https://doi.org/10.26499/li.v39i1.197

[9] Qi W, Sorokina N. Constructing online tourist destination images: a visual discourse analysis of the official Beijing Tourism website[J]. Chinese Semiotic Studies, 2021, 17(3): 421-448. https://doi.org/10.1515/css-2021-2006

[10] Yu Z, Xiao Z, Liu X. A data-driven perspective for sensing urban functional images: Place-based evidence in Hong Kong[J]. Habitat International, 2022, 130: 102707. https://doi.org/10.1016/j.habitatint.2022.102707

[11] Su L, Chen W, Zhou Y, et al. Exploring city image perception in social media big data through deep learning: A case study of Zhongshan City[J]. Sustainability, 2023, 15(4): 3311. https://doi.org/10.3390/su15043311

[12] Xie L, Feng X, Zhang C, et al. Identification of urban functional areas based on the multimodal deep learning fusion of high-resolution remote sensing images and Social Perception Data[J]. Buildings, 2022, 12(5): 556. https://doi.org/10.3390/buildings12050556

[13] Fan R, Li F, Han W, et al. Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16. https://doi.org/10.1109/tgrs.2022.3204345

[14] Zhang F, Salazar-Miranda A, Duarte F, et al. Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery[J]. Annals of the American Association of Geographers, 2024, 114(5): 876-897. https://doi.org/10.1080/24694452.2024.2313515

[15] Huang J, Obracht-Prondzynska H, Kamrowska-Zaluska D, et al. The image of the City on social media: A comparative study using “Big Data” and “Small Data” methods in the Tri-City Region in Poland[J]. Landscape and Urban Planning, 2021, 206: 103977. https://doi.org/10.1016/j.landurbplan.2020.103977

[16] Huang Y, Zheng B. Social media users' visual and emotional preferences of internet-famous sites in urban riverfront public spaces: a case study in Changsha, China[J]. Land, 2024, 13(7): 930. https://doi.org/10.3390/land13070930

[17] Reyes MF, Xie Y, Yuan X, et al. A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 205: 74-97. https://doi.org/10.1016/j.isprsjprs.2023.09.013

[18] Piadyk Y, Rulff J, Brewer E, et al. Streetaware: A high-resolution synchronized multimodal urban scene dataset[J]. Sensors, 2023, 23(7): 3710. https://doi.org/10.3390/s23073710

[19] Chen S, Zang Y, Yang P. City images in transnational travel vlogs from a multimodal perspective: an investigation of 20 port cities worldwide[J]. Online Media and Global Communication, 2024, 3(1): 82-107. https://doi.org/10.1515/omgc-2023-0034

[20] Chen X, Yu J, Zhu Y, et al. Short video-driven deep perception for city imagery[J]. Environment and Planning B: Urban Analytics and City Science, 2024, 51(3): 689-704. https://doi.org/10.1177/23998083231193236

[21] Kang Y, Cho N, Yoon J, et al. Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos[J]. ISPRS International Journal of Geo-Information, 2021, 10(3): 137. https://doi.org/10.3390/ijgi10030137

[22] Xue J, Jiang N, Liang S, et al. Quantifying the spatial homogeneity of urban road networks via graph neural networks[J]. Nature Machine Intelligence, 2022, 4(3): 246-257. https://doi.org/10.1038/s42256-022-00462-y

[23] Liu C, Wang Y, Li W, et al. An urban built environment analysis approach for street view images based on graph convolutional neural networks[J]. Applied Sciences, 2024, 14(5): 2108. https://doi.org/10.3390/app14052108

[24] Rashidan H, Musliman IA, Sani MJ, et al. Semantic labeling of 3D buildings by using graph neural network (GNN)[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2024, 48: 307-311. https://doi.org/10.5194/isprs-archives-xlviii-4-w9-2024-307-2024

[25] Ma X, Zeng T, Zhang M, et al. Street microclimate prediction based on Transformer model and street view image in high-density urban areas[J]. Building and Environment, 2025, 269: 112490. https://doi.org/10.1016/j.buildenv.2024.112490

[26] Wang W, Teng Y, Yan L, et al. Image Experience Prediction for Historic Districts Using a CNN-Transformer Fusion Model[J]. Image Analysis and Stereology, 2025, 44(1): 11-23. https://doi.org/10.5566/ias.3361

[27] Cho M, Kim S, Choi D, et al. Enhanced BLIP-2 Optimization Using LoRA for Generating Dashcam Captions[J]. Applied Sciences, 2025, 15(7): 3712. https://doi.org/10.3390/app15073712

[28] Lee C, Jang J, Lee J. Personalizing text-to-image generation with visual prompts using BLIP-2[J]. 2023.

[29] Huang T, Wang Z, Sheng H, et al. Learning neighborhood representation from multi-modal multi-graph: Image, text, mobility graph and beyond[J]. arXiv preprint, 2021 , 21 (05): 02489.

[30] Zhang J, Chen R, Li S, et al. MGKGR: Multimodal Semantic Fusion for Geographic Knowledge Graph Representation[J]. Algorithms, 2024, 17(12): 593. https://doi.org/10.3390/a17120593

[31] Feng Z, Zeng Z, Guo C, et al. Temporal multimodal graph transformer with global-local alignment for video-text retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(3): 1438-1453. https://doi.org/10.1109/tcsvt.2022.3207910

[32] Hu B, Zhu L, Dong Q, et al. Physiological electrosignal asynchronous acquisition technology: Insight and perspectives[J]. IEEE Transactions on Computational Social Systems, 2024, 11(1): 5-24. https://doi.org/10.1109/tcss.2024.3350958

[33] Han Y. The Study on the Multimodal Discourse Construction and Communication of Zhengzhou’s National Central City Image[J]. Journal of Business and Management Studies, 2025, 7(1): 222-226. https://doi.org/10.32996/jbms.2025.7.1.17

[34] Guan S, Cheng X, Bai L, et al. What is event knowledge graph: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7): 7569-7589.

[35] Joshi R. Introduction to graph neural network: A systematic review of trends, methods, and applications[J]. Applied Graph Data Science, 2025: 1-16. https://doi.org/10.1016/b978-0-443-29654-3.00017-x

[36] Yang Z, Zhang J, Li Z. Multi-scale time series prediction model based on deep learning and its application[J]. PLoS One, 2025, 20(7): e0325474. https://doi.org/10.1371/journal.pone.0325474

[37] Lisena P, Meroño-Peñuela A, Troncy R. MIDI2vec: Learning MIDI embeddings for reliable prediction of symbolic music metadata[J]. Semantic Web, 2022, 13(3): 357-377.

https://doi.org/10.3233/sw-210446
[38] Biometric Recognition: 18th Chinese Conference, CCBR 2024, Nanjing, China, November 22–24, 2024, Proceedings, Part II[M]. Springer Nature, 2025. https://doi.org/10.1007/978-981-96-1071-6_9
[39] Gu X, Liu C, Wang S. Biometric Recognition[J]. Lecture Notes in Computer Science, 2013, 8232: 34-42.

English Composition Topic and Emotion Classification Based on LDA-BERT Fusion and Dual GRU model

Lubing Shang
Anyang Vocational and Technical College, Anyang, Henan, 455000, China
E-mail: Lubing_Shang@outlook.com

Keywords: LDA model, BERT, GRU, theme recognition, sentiment polarity classification

Received: September 5,2025

With the development of globalization and artificial intelligence, the traditional manual correction of English compositions has problems, such as low efficiency and a lack of standardized evaluation. Therefore, there is an urgent need for automated analysis. The study aims to develop a precise model for identifying English essay themes and classifying emotional perspectives. Latent Dirichlet allocation is used to combine Probase prior knowledge and BERT fusion technology for topic modeling, and a Bi-GRU network with an attention mechanism is used for emotion classification. The experiment was carried out based on a dataset of 66,000 English compositions, including TOEFL, IELTS, and classroom compositions, and verified the effectiveness and superiority of the joint model in topic identification and emotion classification tasks. The experiment showed that the topic recognition model had a perplexity of 803, coherence of 0.831, and inter-class distance of 3.2 when there were 25 topics. When the threshold of the emotion classification model was 0.45, the accuracy reached 80.1% and the F1 value was 73.7%. The F1 value of 2,000 tests was 76.1%, which was better than the comparison model. Research has shown that this solution effectively solves problems such as insufficient semantic understanding in existing technologies, provides scientific support for intelligent evaluation systems, and promotes the intelligent development of educational evaluation.

Povzetek: Raziskava predstavi skupni model za samodejno prepoznavanje tem in čustvenih perspektiv v angleških esejih, ki z združitvijo LDA z BERT ter Bi-GRU z mehanizmom pozornosti dosega boljšo natančnost in semantično razumevanje pri inteligentnem vrednotenju pisnih nalog.

1 Introduction

With the deep integration of Artificial Intelligence (AI) and education, intelligent English composition evaluation systems have gradually become an important tool for assisting teaching [1]. However, the existing evaluation system has obvious shortcomings in the analysis from the emotional perspective. These systems cannot accurately grasp the emotional tendencies of students' compositions based on specific themes, resulting in a lack of depth and pertinence in the evaluation [2, 3]. Therefore, developing algorithms that can accurately identify the themes, emotions, and viewpoints of English composition sentences is of great significance for improving the quality of intelligent review and assisting teachers in comprehensively evaluating students' writing skills. Currently, advanced technologies in text sentiment analysis mainly rely on deep learning models, which perform well in general text sentiment classification tasks [4, 5]. Some researchers have attempted to incorporate attention mechanisms into the model to enhance its ability to capture key emotional information [6]. However, in the context of sentiment analysis in English compositions, most existing models overlook the guiding role of themes in emotional judgment, making it difficult to distinguish the emotional differences of the same expression under different themes. At the same time, there is a lack of refined modeling of the correlation between vocabulary and topic within sentences, resulting in a disconnect between sentiment analysis results and the core ideas of the composition [7]. Therefore, this study proposes a novel algorithm that integrates topic embedding and bidirectional neural networks. This algorithm deeply integrates the topic word embedding with the sentence word vector and utilizes Bidirectional Gated Recurrent Unit (Bi-GRU) to capture contextual information bidirectionally. It combines a topic-level attention mechanism to dynamically allocate vocabulary weights, ultimately achieving topic-based prediction of sentence sentiment probability distribution.

The innovation of the research lies in proposing the joint optimization of Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT). By preprocessing with Probase knowledge base and BERT deep semantic encoding, combined with neural encoding and decoding architecture, the topic recognition ability is enhanced. As an exception, a Bi-GRU sentiment classifier is synchronously designed, integrating topic attention mechanism to enhance contextual semantic modeling. This study adopts a dual model collaborative mechanism to achieve topic-oriented sentiment probability prediction through topic embedding and bidirectional network fusion, effectively breaking

through the limitations of traditional model semantic representation. This study aims to integrate the knowledge-enhanced topic model of Probase and BERT, combined with the topic-guided attention mechanism. It effectively solves the problems of shallow semantic understanding and separation of topic and emotion in traditional English composition analysis methods.

2 Related works

The current English composition topic recognition and Sentiment Polarity Classification (SPC) still faces problems such as blurred topic boundaries, misjudgment of emotional polarity, and insufficient understanding of context. Zhu proposed the Dirichlet process modeling+quantitative word frequency matrix method to efficiently utilize large amounts of text data in English curriculum development to extract key teaching themes. This method performed excellently in accuracy, recall, and precision [8]. Lin et al. proposed a probabilistic convolutional kernel implementation method based on memristors and a Bayesian forward training method to address the challenges faced by Bayesian Convolutional Neural Networks (CNNs) in hardware implementation. On the MNIST and CIFAR10 datasets, this method achieved accuracies of 98.67% and 87.81%, and achieved excellent out of distribution detection performance [9]. Cochran et al. proposed an identification method based on BERT and its variants to address the issue of causal relationship recognition and feedback automation in scientific explanation writing. The model could effectively identify the existence of causal relationships and was positively correlated with writing quality, but due to imbalanced data categories, there were still limitations in determining specific causal relationships, and further optimization was needed [10]. Cho et al. proposed a dual-scale BERT self-attention transformer CNN joint model to address the problem of insufficient modeling of overall and specific feature evaluation in automatic paper grading. This model combined dual-scale encoding with a multi-task learning framework. On the ASAP++ and TOEFL 11 datasets, this method improved the overall score by 2.0% and the multi-feature average performance by 3.6% compared to the single-task model [11]. Lagutina et al. proposed a vector model based on character, word, and syntactic features to automatically classify the language proficiency of short texts according to the Common European Framework of Reference for Languages. The F1 value of the support vector machine on the English corpus was 67%, while the basic version of the BERT computer-aided pattern evaluation and recognition joint method had the best performance in the BERT variant, with an F1 value of 69%. This indicated that the method had practical application potential [12].

Sharma et al. proposed a method that combines BERT pre-trained word embedding with Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) models to address the challenges of language understanding and feature expression in automatic essay scoring. The average quadratic weighted kappa coefficient of the SVR model reached 0.81, indicating its significant advantages in improving the accuracy and robustness of automatic scoring [13]. Rogers proposed an automatic optimization method based on a genetic algorithm to address the problem of difficult fair comparison between deep neural network architecture and hyperparameters in sentence classification tasks. The use of F1 value evolution models generally outperformed accuracy-based models, verifying the effectiveness in improving generalization ability and optimizing architecture selection [14]. Liu et al. proposed a deep learning model that combines residual networks and GRU to address the difficulties in feature extraction and limitations of traditional methods in the recognition of large and complex electrical objects. The average recognition rate was 95.56%, and it had good noise resistance [15]. Prakash et al. proposed a model based on CNN and GRU to address the difficulties in feature extraction and insufficient emotion discrimination in automatic speech emotion recognition. This method showed high recognition accuracy in three emotion state classification tasks, indicating its good robustness and generalization ability in speech emotion recognition applications [16]. Behrouzi et al. proposed a multimodal deep learning model based on GateGRU to address the complex classification of movie trailers and the limited performance of traditional models. This method significantly outperformed existing advanced models in classification performance, improving the accuracy and efficiency of multi-label movie type recognition [17].

In summary, the existing English composition topic recognition and SPC technology have problems such as shallow semantic understanding and incomplete feature extraction, making it difficult to accurately capture complex semantics and emotional tendencies. The LDA model and BERT topic recognition model have good semantic representation capabilities, while GRU utilizes Bi-GRU to enhance the capture of contextual information in text. Therefore, this study proposes an optimized LDA fusion BERT topic recognition model and an based on probase Bi-GRU SPC model. The collaboration between the two provides an efficient and accurate solution for English composition analysis.

Table 1: Comparison of research on existing methods

References	Research Focus	Methods Used	Dataset/Task	Key Indicators	Main Limitations / Differences from This Study
Text mining and topic modelling in English teaching: Extracting key themes and concepts for effective curriculum development.	Course Topic Extraction	Dirichlet Process Model	English Teaching Text	Accuracy, Recall, Precision	The association between topics and emotions is not considered, focusing on macro topic discovery.
Using BERT to identify causal	Causal	BERT and Its	Scientific Explanation	Causal	There is a problem of

structure in students' scientific explanations.	Relationship Recognition	Variants	Writing	Relationship Recognition Accuracy	category imbalance, and domain knowledge is not integrated for semantic enhancement.
Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading.	Automatic Essay Scoring	Dual-scale BERT-Self-Attention-CNN	ASAP++, TOEFL11	Overall Score Improvement by 2.0%	It focuses on overall scoring and does not deeply analyze the fine-grained association between topic structure and sentiment polarities.
Text Classification by CEFR Levels Using Machine Learning Methods and the BERT Language Model.	Text Difficulty Classification	SVM, BERT Variants	CEFR Standard English Corpus	F1 value (69%)	It focuses on language proficiency and does not involve the combination of topic modeling and sentiment analysis.
Modeling essay grading with pre-trained BERT features.	Automatic Essay Scoring	BERT + SVR/LSTM	Standard Essay Scoring Dataset	Weighted Kappa (0.81)	The model focuses on score prediction and lacks an explanation of the interactive mechanism between the internal topics and emotions of compositions.
A comparative analysis of deep neural network architectures for sentence classification using genetic algorithm.	Sentence Classification	Genetic Algorithm Optimized DNN	Standard Sentence Classification Task	F1 value	It focuses on the optimization of general architectures and does not design for the topic-emotion characteristics in the field of composition.

3 Methods and materials

3.1 Topic recognition model based on improved LDA and BERT feature fusion

With the acceleration of globalization and the development of AI technology, English writing ability has become a key skill for cross-cultural communication and academic research. In English composition evaluation, topic recognition is an important dimension for determining the relevance of content, and traditional LDA models are difficult to meet the evaluation needs of English compositions [18, 19]. Therefore, this study proposes an LDA topic recognition model that integrates BERT optimization. Before topic recognition, it is necessary to preprocess the essays to be submitted, as shown in Figure 1.

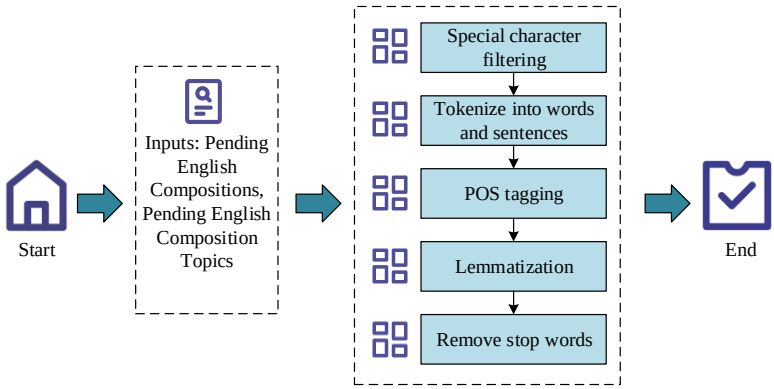


Figure 1: Preprocessing flowchart for essays awaiting approval

In Figure 1, the model needs to first read in the pending English essay and its title, and then filter out special characters using a special character set. Then, it uses the CoreNLP toolkit for word segmentation and sentence segmentation, and then performs part of speech tagging on the vocabulary after word segmentation and sentence segmentation. By combining the maximum entropy model with context to determine the part of speech, the different parts of speech annotated by vocabulary in different contexts are distinguished. Then, the vocabulary is restored to the dictionary prototype. Finally, stop words such as auxiliary verbs and prepositions are removed to obtain a collection of word roots that can represent the composition, laying the foundation for subsequent topic recognition and sentiment analysis [20, 21]. According to the root set obtained from the preprocessing module, the probability distribution of concept a in English composition is obtained, as shown in equation (1)

$$f(a_l|\mathbb{C}^{(t)}) = \frac{p(\mathbb{C}^{(t)}|a_l)\sum_{c_j^{(t)}\in a_l} n(a_l, c_j^{(t)})}{p(\mathbb{C}^{(t)})} \propto \sum_{c_j^{(t)}\in a_l} n(a_l, c_j^{(t)}) \prod_{j=1}^{N} p(c_j^{(t)}|a_l) \quad (1)$$

In equation (1), f is the posterior probability of the concept. C is the collection of pre-processed word roots in the composition. t is the t-th essay, a_l is the l-th concept, j is the root, and N is the total number of root sets. In the filtered set, the top three concepts in the probability distribution of the concept set of each English composition will be included in the global concept set of the corpus. Then, duplicate concepts in the set are removed, and concept clusters of corresponding themes are formed through concept clustering. The k-medoids algorithm is used for concept clustering, and cosine distance is used as the similarity measure. Through silhouette coefficient analysis, the optimal number of clusters is determined to be 8. Probase concepts are filtered based on word frequency and typicality scores, and the top 15% concepts with scores greater than 0.7 are retained to obtain concept space. Therefore, this study will integrate the optimized LDA and BERT pre-trained language model to construct an optimized LDA-fused BERT topic recognition model (LDA-BERT model, PL-BERT), as shown in Figure 2.

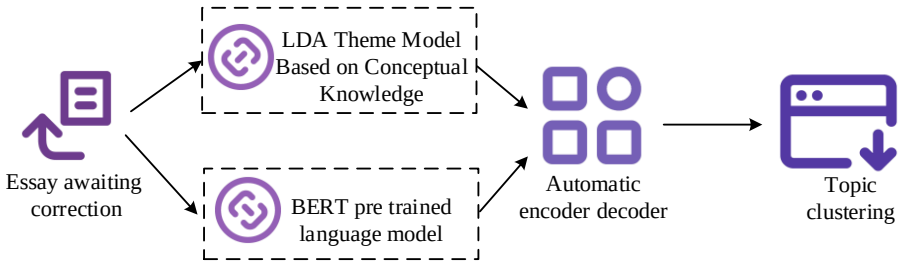


Figure 2: Topic recognition model based on optimized LDA fusion BERT (Source: https://iconpark.oceanengine.com)

According to Figure 2, the preprocessed English essay is first input into the LDA topic model based on the Probase concept knowledge base, and then a document concept set is generated using Probase. After filtering and clustering, it is used as an LDA prior parameter to obtain a topic vector that reflects global semantics. At the same time, the composition is processed by the BERT model, input in a special format, and output as a local semantic vector containing contextual syntactic information using the last layer of special classification tags. Then, the global topic vector and the local semantic vector are directly spliced, and the global topic distribution of LDA and the local context semantics of BERT are combined. The fusion vector L is shown in equation (2).

$$L = \text{gram}\{e_1, e_{2\ldots} e_B\} + \{f_1, f_{2\ldots} f_N\} \quad (2)$$

In equation (2), B is the number of topics, e is the essay topic vector, f is the essay semantic vector, and N is the number of semantics. The fusion vector is input into an automatic encoding decoding neural network topic model based on Gaussian distribution. The encoding layer generates topic vectors through multi-layer perceptrons, which are then transformed into topic probability distributions using softmax. The output latent topic vector R is shown in equation (3).

$$R = st\max(Av) \quad (3)$$

In equation (3), A is a linear transformation matrix, v is a vector with a Gaussian distribution, and st is a softmax function. The decoding layer reconstructs document semantics through potential topic vectors to generate documents, as shown in equation (4).

$$\log c\left(d_{r,i} \middle| \beta, b^r\right) = \log\left(b^r \cdot \beta\right) \quad (4)$$

In equation (4), d is location search, r is document, i is word, and β is topic vocabulary distribution matrix. b^r represents the topic distribution vector of the document. r and c are probabilities. The next step is to cluster each topic embedding representation using k-means clustering to obtain deep latent topic representations. Based on the above model structure, an optimized LDA fusion BERT based topic recognition algorithm can be obtained, as shown in Figure 3.

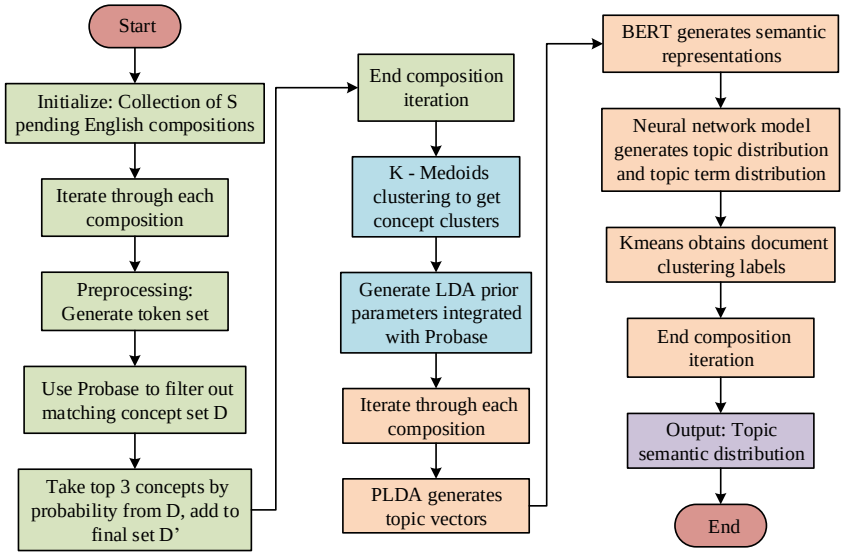


Figure 3: Theme recognition algorithm based on optimized LDA fusion BERT

In Figure 3, the BERT fusion algorithm first counts the number of articles, preprocesses each article, and then outputs a set of root words [22]. Next, the Probase concept knowledge base is utilized to generate a set of concepts that conform to the English composition corpus, and the top three probabilistic concepts are selected to form the set. The prior parameters of the Probase LDA Model (P-LDA) are obtained through clustering. Then, each essay uses the P-LDA model to obtain topic vectors and the BERT model to obtain contextual semantic vectors. The two are proportionally concatenated and input into a neural network topic model, which generates topic and term distributions through an automatic encoding-decoding structure. Finally, the topic tags are obtained through the k-means clustering algorithm, which enables in-depth mining of the hidden topic semantics of English compositions. The accuracy of topic identification and the completeness of semantic representation are significantly improved.

3.2 SPC model based on GRU network

Emotional perspective analysis, as an important dimension in English composition evaluation, directly affects the comprehensive evaluation of students' writing ability based on its accuracy. However, existing sentiment analysis models have significant shortcomings when dealing with complex texts such as English compositions [23, 24]. Bi-GRU can effectively extract long sequence text information. Therefore, this study proposes a Bi-GRU SPC model based on a topic-level attention mechanism, whose architecture is shown in Figure 4.

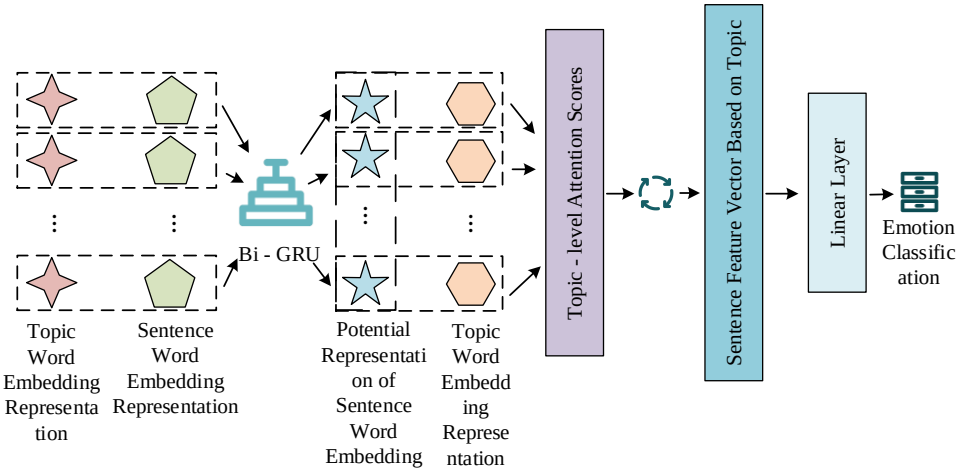


Figure 4: Schematic diagram of the optimized Bi-GRU SPC model

In Figure 4, the word embedding layer concatenates the sentence word embeddings generated by the Word2Vec model with the latent topic vectors of the pending essay to form the model input. The next step is to process the input sequence using Bi-GRU to obtain the contextual semantic latent representation of each term. The topic-level attention mechanism layer calculates the attention scores of each vocabulary in the sentence based

on the topic embedding vector, dynamically adjusts the weights, and weights the sum to focus on the key features that express the topic's emotions. Among them, the theme embedding vector is shown in equation (5).

$$T_o = \tanh(q_s y_o + p_s) \tag{5}$$

In equation (5), q_s is a $d_e \times d_h$ matrix. y_o and q_s are all of d dimensions. T_o is the output vector, q_s is the weight matrix, y_o is the input hidden state vector, and p_s is the bias vector. The calculation of the attention score for each vocabulary on the topic is shown in equation (6).

$$k_o = \frac{\exp(u_o \cdot h)}{\sum_{j=1}^{m} \exp(u_j \cdot h)}, \sum_{o=1}^{m} k_o = 1 \tag{6}$$

In equation (6), u is a d_h-dimensional vector, h is of d_h dimension, and k outputs a 1-dimensional value. k is the attention weight, u_o represents the intermediate semantic vector of the o -th word, and ex is the transformation function. j represents the index of the j -th word, h is the queryable vector, and m is the total number of elements involved in attention calculation. The latent semantics of the sentence are shown in equation (7).

$$d = \sum_{i=1}^{s} k_i y_o \tag{7}$$

In equation (7), d is of d_h dimension. d is the final eigenvector obtained by weighted summation, as shown in equation (8).

$$z = soft\max(sigmoid(Y_s, s)) \tag{8}$$

In equation (8), Y is a $d_out \times d_h$ matrix. s is the d_h dimension, and z is the d_out dimension. The output dimension of the activation function is the same as its input dimension, and no Dropout layer is used. z is the final probability of outputting sentiment classification, s is the latent semantics of the sentence, and Y is the change matrix. This architecture provides key criteria for evaluating the quality of essay themes and viewpoints through inter layer collaboration. Based on the above architecture, this study has improved the topic-based sentiment classification algorithm, as shown in Figure 5.

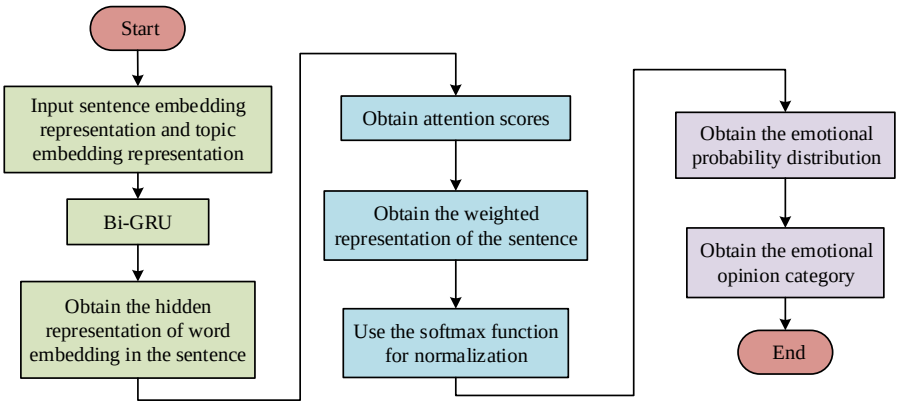


Figure 5: Flowchart of theme-based emotion classification algorithm

In Figure 5, the algorithm first inputs the word embedding representation set and the topic word embedding representation of the English composition sentence to be approved. Next, the topic word embedding and sentence word embedding are concatenated and input together into the Bi-GRU to obtain the hidden representation of the sentence word embedding. Subsequently, the topic-level attention mechanism is used to process the hidden representation set, and the topic vector of the sentence is used as the "query" benchmark. The correlation between the hidden state of each word in the sentence and the topic vector is calculated to obtain the attention score. Based on the attention score, the hidden representation set is weighted and summed to form a sentence representation that can dynamically focus on topic-related emotional words. Then, the weighted representation is processed through linear layers and softmax functions to achieve the probability distribution of emotions related to the topic. Finally, the tagging process ends and returns the sentiment probability distribution, which is used to present the results of sentiment analysis based on the topic of the sentence. To analyze the effectiveness of the themes and sentiment polarities obtained from relevant models, this study designs an English composition theme viewpoint quality evaluation method. This method is used to score the themes and sentiment polarities of the composition, as shown in Figure 6.

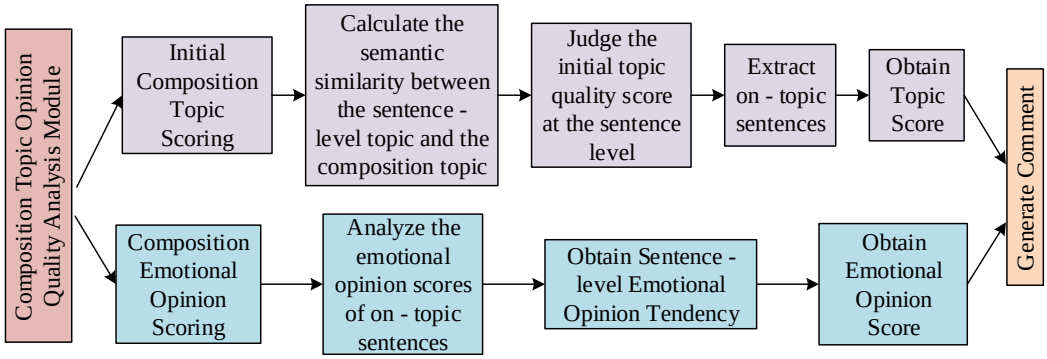


Figure 6: Quality evaluation method for English composition theme viewpoints

In Figure 6, in the topic relevance path, step 1 is to extract the semantic vectors of the sentence and the essay topic through the BERT model, and use the cosine similarity algorithm to calculate the semantic matching degree between the two. The scoring mechanism is constructed based on this similarity to quantify the sentence-level topic quality score, achieve accurate recognition of the target text, select the target sentence, and ultimately obtain the topic score. The sentiment polarity path first analyzes the sentiment polarity scores of relevant sentences, delves into the emotional tendencies of related content, and then obtains sentence-level sentiment polarity scores. After the theme rating and sentiment polarity rating are determined, the system will generate comments and integrate these indicators to produce a comprehensive and personalized evaluation. This method constructs a dual-focus evaluation system, ensuring a close fit between the composition and the theme, while taking into account the richness and relevance of the emotional perspective. This provides a comprehensive evaluation perspective for the quality of English composition topic perspectives.

4 Results

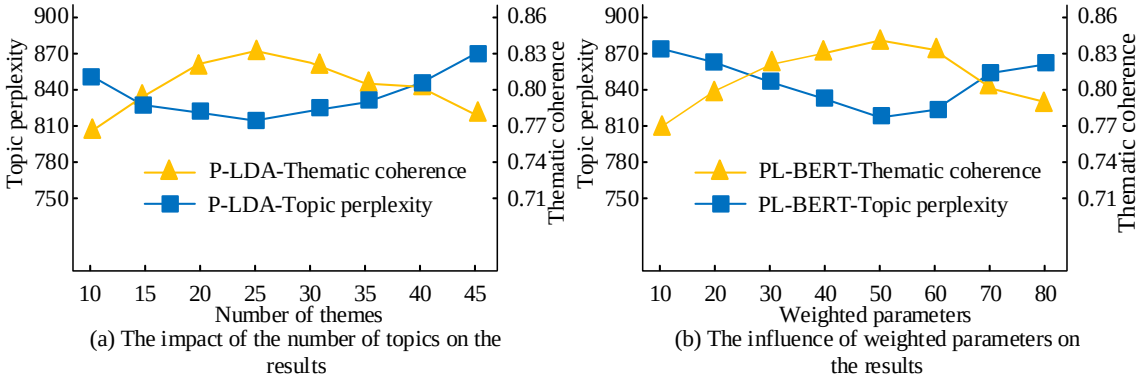
4.1 Topic recognition model test results

To conduct experimental testing of the designed model, the study uses 40,000 EFCAMDAT, 16,000 TOEFL11 Corpus, and 10,000 IELTS Essay datasets, totaling 66,000 English essays. The study divides 66,000 essays into training sets and test sets at a ratio of 8:2. The model is trained using the Adam optimizer, the learning rate is set to 5e-5, the batch size is 32, and a total of 50 rounds are trained. An early stopping method is used to prevent overfitting. This study also builds a simulation testing platform to test the performance of the model. The platform is tested with SpaCy 3.5.0 and processes with Pandas 2.1.0. The detailed information is shown in Table 2.

Table 2: Detailed information of the testing platform

Hardware Information		Software information	
Name	Type	Name	Type
CPU	AMD Threadripper 3990X	Text preprocessing tool	NLTK 3.8.1
GPU	NVIDIA A100	Natural Language Processing Library	SpaCy 3.5.0
Storage device	1TB NVMe SSD+4TB SATA SSD	Data processing tools	Pandas 2.1.0
Dynamic Calculation Chart Tool	PyTorch 2.0	Virtual Environment Tools	Anaconda 2023.07
Static Calculation Chart Tool	Pandas 2.1.0	Development tool	PyCharm 2023.3

All comparative experiments are independently repeated five times under identical conditions. The final reported results are averaged, and key indicators are tested with two-sided t-tests. To test the performance of PL-BERT in topic recognition of English compositions, this study uses topic perplexity and topic coherence as judgment indicators. The experimental comparison between the number of topics in P-LDA and the vector weighting parameters in PL-BERT is shown in Figure 7.

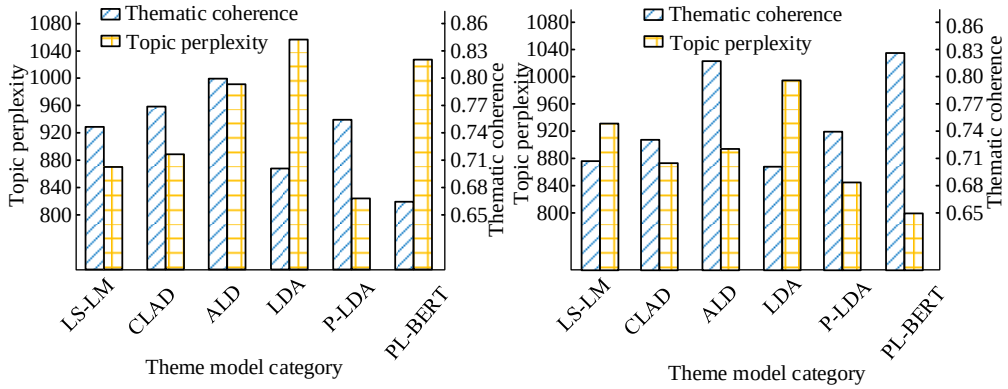


(a) The impact of the number of topics on the results

(b) The influence of weighted parameters on the results

Figure 7: The impact of the number of themes and weighting parameters on the model

In Figure 7 (a), as the number of topics gradually increases from 10, the topic perplexity of P-LDA gradually decreases from 857. When the number of topics is 25, the topic perplexity drops to the lowest, which is 803. As the number of topics increases, the level of thematic confusion also gradually increases. The thematic coherence reaches its peak when the number of themes is 25, with a thematic coherence of 0.831. Therefore, under the experimental data, when the number of topics is 25, the model has the best clustering effect on English essay topics. In Figure 7 (b), when the weighting parameter is 50, the topic perplexity is the lowest, at 821. At this point, the coherence of the theme reaches its highest value of 0.864, indicating that when the weighting parameter is 50, PL-BERT can achieve optimal performance. When the number of topics deviates from 25 or the weighted parameters deviate from 50, both indicators fluctuate, which also verifies the rationality of the parameter settings. The metric values at parameter 25 and parameter 50 show statistically significant differences from adjacent parameters (p<0.05). The proposed PL-BERT is compared with Large-Scale Language Models (LS-LM) [18], Computer Language-Aided models (CLAD) [19], Automatic Language Detection (ALD) [20], LDA, and P-LDA models, as shown in Figure 8.



(a) Evaluation results of different models on daily classroom composition

(b) Evaluation results of different models on TOEFL and IELTS essay writingdata

Figure 8: Differences in model influence under different datasets

In Figure 8 (a), in the daily classroom composition dataset, the theme perplexity of PL-BERT is about 821, with a theme coherence of 0.823, while the perplexity of P-LDA is about 824, with a coherence of about 0.754. In the experiment, the topic perplexity of LS-LM, CLAD, and ALD models is higher than that of PL-BERT, and the topic coherence is lower than that of PL-BERT. In Figure 8 (b), in the TOEFL and IELTS essay datasets, PL-BERT has a perplexity level of approximately 803 and a coherence level of approximately 0.831. Its performance in testing topic perplexity and coherence is still superior to other models. Whether in the daily classroom composition dataset or the TOEFL/IELTS composition dataset, PL-BERT exhibits good performance advantages in different language usage contexts. This also validates the effectiveness and robustness of the model fusion strategy on different datasets. The PL-BERT model outperforms all baseline models by a statistically significant margin (p<0.01) on both datasets. To verify the theme clustering effect of PL-BERT, this study uses TOEFL/IELTS essays as experimental data. The experiment uses inter-class distance and intra-class compactness as evaluation indicators, selects four similar topics (dividing into 8 vocabulary words), and conducts visual clustering analysis based on 2D UMAP technology [21]. The traditional LDA model is used as the comparison benchmark, as shown in Figure 9.

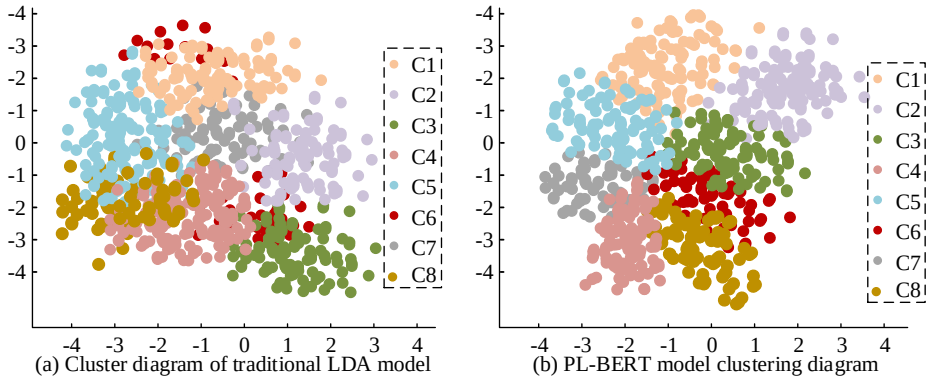


Figure 9: Clustering effect of different models (Image source: self drawn by the author)

In Figure 9 (a), the clustering points of LDA are loosely distributed, with an inter-class distance of only 1.8 and an intra-class compactness of 0.68. The silhouette coefficient of the traditional LDA model is only 0.28. Different topic clusters overlap, and some topic clusters exhibit obvious separation. Among the distribution characteristics of topic clusters, LDA has obvious shortcomings in classifying topic clusters. In Figure 9 (b), the PL-BERT clustering points exhibit clear inter-class separation and high intra-class clustering. The distance between topic clusters under PL-BERT is 3.2, and the compactness of samples within the same topic is 0.45. The contour coefficient of the PL-BERT model has been increased to 0.62. Compared with the inter-class distance index and intra-class compactness of the two models, PL-BERT has a higher inter-class distance, clear differentiation, lower intra-class compactness, and clear division of topic samples. This validates the advantages of the model in semantic space mapping and topic clustering. The PL-BERT model outperforms the LDA model significantly in both inter-class distance and intra-class compactness (p<0.001).

4.2 Opinion classification test results

The study conducts experiments using data from the Common European Language Reference Framework (CEFR) at levels B1 to C1 in the test set. To verify the experimental effectiveness of the emotional perspective analysis model, this study tests and evaluates the emotional perspective attitude data extracted from English compositions on various topics. This study uses precision, recall, and F1 value as evaluation indicators to conduct comparative experiments between the optimized Bi-GRU SPC model and the topic-based sentiment classification model, as shown in Figure 10.

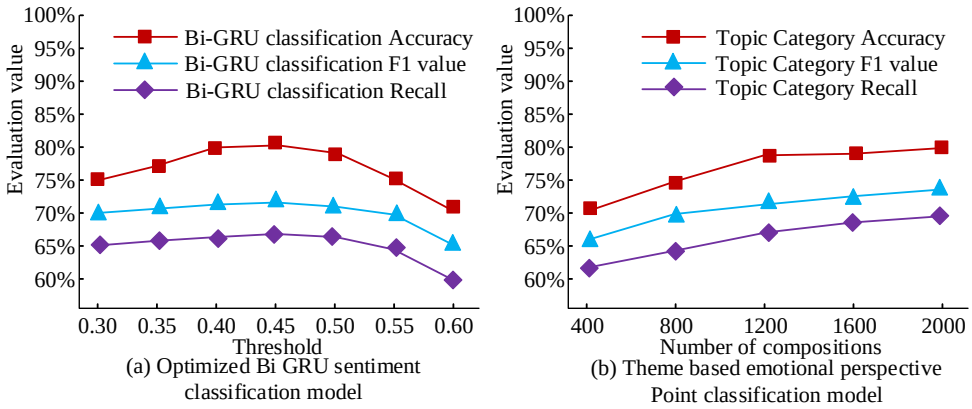


Figure 10: Comparison of model index testing under different conditions

In Figure 10 (a), the optimal accuracy, recall, and F1 value of the optimized Bi-GRU SPC model can be tested by adjusting the threshold. As the threshold increases, each indicator shows a trend of first increasing and then decreasing. When the threshold reaches 0.45 (p<0.05), the accuracy, recall, and F1 value of the model reach 80.1%, 68.2%, and 73.7%, demonstrating the best performance in predicting positive cases and capturing positive cases in experiments. In Figure 10 (b), as the number of essays increases from 400 to 2,000, the index indicators of optimal accuracy, recall, and F1 value all show a gradually increasing trend. However, within the range of 400-1,200 essays, the growth rate of each indicator value is significantly faster than that of the following 1,200-2,000. However, this experiment has not yet tested the limit on the number of articles with peak values for each indicator, demonstrating that the model has good compatibility with data capacity and stable robustness. To verify the performance advantages of the theme SPC model, this study selects the Based on the Recurrent Neural Network

Classification Model (BCNNC) [22], Evidence is Stacked on the GUR Neural Network (E-GUR) [23], and Neural Network Fusion Model (NNF) [24] as baselines. This experiment compares accuracy, precision, recall, and F1 value indicators. The test results are shown in Figure 11.

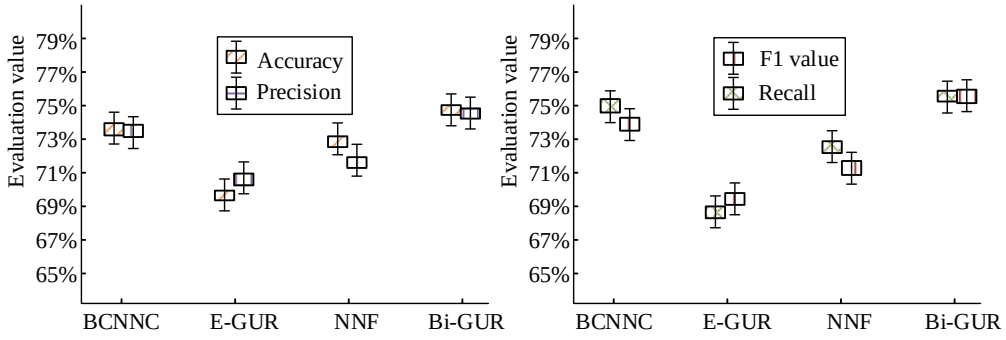


Figure 11: Comparison of indicators for different models (Image source: self drawn by the author)

In Figure 11 (a), among all the models involved in the experiment, the precision and accuracy of Bi-GUR are 75.9% and 75.4%, slightly higher than the 73.5% and 73.8% of BCNNC. The precision and accuracy of E-GUR are the lowest, at 70.8% and 70.1%. Bi-GUR, which has stronger comprehensive performance value and environmental adaptability, performs more outstandingly. In Figure 11 (b), the recall rate and F1 value of Bi-GUR are 76.3% and 76.1%. The E-GUR with the worst performance in this part of the experiment is 69.8% and 70.6%, which are 5.5% lower than the corresponding indicators of Bi-GUR. Further comparison with other models shows that Bi-GUR has a recall index 4.2% higher than the theme level NNF, and an F1 value 0.8% higher than BCNNC. Overall, Bi-GUR outperforms most comparison models in all indicators and demonstrates significant advantages in English composition SPC tasks. Especially in terms of recall rate and F1 value, it demonstrates the effectiveness of capturing contextual semantic dependencies through Bi-GRU and enhancing emotional feature extraction through topic-level attention mechanism (p<0.05). Bi-GUR can more comprehensively identify sentiment polarity information in essays.

5 Discussion

The LDA-BERT and Bi-GRU integrated model demonstrates exceptional performance in the tasks of topic identification and sentiment classification of English essays. Its core advantage stems from the synergistic innovation in its architecture. This study achieves deep semantic association by constructing a dual-model framework. In terms of topic identification, it innovatively incorporates the Probase concept knowledge base as a semantic prior for LDA, effectively addressing the data sparsity issue of traditional LDA models in short text processing. Simultaneously, deep integration with the BERT model leverages its powerful contextual semantic understanding capability to compensate for LDA's shortcomings in handling polysemous words and complex syntactic structures. The latent topic vectors generated by the neural encoder after the fusion of these two components possess both global topic distribution and local semantic details, significantly enhancing topic identification performance. For sentiment classification, a Bi-GRU model based on a topic-level attention mechanism is designed. By using topic embedding vectors extracted from PL-BERT as queries for the attention mechanism, the model dynamically focuses on emotional vocabulary highly relevant to the core topic. Compared to existing advanced methods, the fundamental breakthrough of this research lies in achieving a deep association between topics and sentiments. The Bi-GRU model performs well in terms of overall article scoring, but it ignores fine-grained topic-emotion correlations, and the BERT+SVR/LSTM combination lacks explicit modeling of the topic-emotion interaction mechanism within the article. In contrast, this study achieves a performance breakthrough through the following mechanism. The introduction of Probase prior knowledge guides the model to focus on semantically clear concept clusters, significantly increasing topic coherence to 0.831. The deep semantic fusion of BERT accurately disambiguates polysemous words and complex sentence structures, reducing model perplexity to 803. The topic-level attention mechanism prioritizes emotional expressions highly relevant to the core topic, significantly improving recall while maintaining a precision of 80.1%, ultimately achieving an F1 value of 76.1%. However, the study still has limitations, including insufficient cross-linguistic generalization capability, limited robustness to noisy input, and difficulties in understanding complex rhetorical devices such as metaphors and irony. Future work should focus on lightweight design, cross-language transfer learning, multi-modal fusion, dynamic update mechanism, etc., to enhance the applicability and scalability of the model under different load conditions. This promotes the development of intelligent English composition evaluation in a more accurate and efficient direction.

6 Conclusion

In the context of intelligent education, traditional English composition evaluation urgently needs to innovate with automated analysis technology due to its efficiency and objectivity deficiencies. This study focuses on topic recognition and emotion classification, and constructs a dual model framework of "optimized LDA fusion BERT" and "optimized Bi-GRU". The topic recognition model achieves deep topic extraction through Probase concept screening, BERT semantic vector fusion with LDA prior parameters, and neural network encoding and decoding. The sentiment classification model utilizes Bi-GRU and a topic-level attention mechanism to dynamically capture text sentiment dependencies. Experimental verification showed that the topic model performed outstandingly on the TOEFL/IELTS dataset. When the number of topics was 25, the perplexity was 803, the coherence was 0.831, the inter-cluster distance was 3.2, and the intra-class compactness was 0.45, which was better than the traditional LDA model. The advantages of semantic space mapping and topic clustering were significant. At a threshold of 0.45, the emotion model had an accuracy of 80.1% and an F1 value of 73.7%. The F1 value in the 2,000-essay test increased to 76.1%, which was 4.2% higher than the recall rate of BB-TWTR. The robustness was stable under different data capacities. The model designed in this study breaks through the limitations of the semantic understanding of traditional methods by strengthening semantic representation and topic association. However, its architecture is designed for English and lacks cross-language generalization capabilities. It has limited robustness to noisy input containing spelling errors and non-standard expressions. Moreover, it still has obvious deficiencies in understanding complex rhetorical techniques such as metaphor and irony. Future work can enrich the emotional dimension with multi-modal features by introducing cross-language transfer learning, and integrate a dynamic word vector update mechanism that adapts to language evolution. It can also expand application scenarios and promote the development of intelligent English composition evaluation in a more accurate and diversified direction.

References

[1] Istiqomah A A, Sari C A, Susanto A, et al. Facial Expression Recognition using Convolutional Neural Networks with Transfer Learning Resnet-50. Journal of Applied Informatics and Computing, 2024, 8(2): 257-264. DOI: 10.30871/jaic.v8i2.8329.

[2] Wadawadagi R, Tiwari S, pagi V. Polarity-aware deep attention network for aspect-based sentiment analysis. Progress in Artificial Intelligence, 2025, 14(1): 33-48. DOI: 10.1007/s13748-024-00352-x.

[3] Kristina K, Suarjaya I M A D, Wiranatha A A K A C. Stock Sentiment Prediction of LQ-45 Based on News Articles Using LSTM. Journal of Applied Informatics and Computing, 2025, 9(4): 1154-1162. DOI: 10.30871/jaic.v9i5.10157.

[4] Jones C, Roulet V, Harchaoui Z. Revisiting Convolutional Neural Networks from the Viewpoint of Kernel-Based Methods. Journal of Computational and Graphical Statistics, 2023, 32(4): 1237-1247. DOI:10.1080/10618600.2022.2163649.

[5] Xue M L, Wang D Y. Cross-modal sentiment analysis on social media using improved nonverbal representation learning and GHRNN fusion. Informatica, 2025, 49: 35-50. DOI: https://doi.org/10.31449/inf.v49i34.9130.

[6] Wang S, Li F. Deep Reinforcement Learning with Convolutional Neural Networks for Optimizing Supply Chain Inventory Management. Informatica, 2025, 49: 201-222. DOI: https://doi.org/10.31449/inf.v49i26.8396.

[7] Shi J, Li Z, Lai W, LI F, Shi R, Feng Y, et al. Two end-to-end quantum-inspired deep neural networks for text classification. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 4335-4345. DOI:10.1109/TKDE.2021.3130598.

[8] Zhu J. Text mining and topic modelling in English teaching: Extracting key themes and concepts for effective curriculum development. Journal of Computational Methods in Sciences and Engineering, 2025, 25(2): 1210-1222. DOI: 10.1177/14727978241298468.

[9] Lin Y, Zhang Q, Gao B, Tang J, Zhao H, Qin Q, et al. High-Efficient Memristor-Based Bayesian Convolutional Neural Networks for Out-of-Distribution Detection by Uncertainty Estimation. IEEE Transactions on Electron Devices, 2025, 72(1): 206-214.DOI:10.1109/TED.2024.3497917.

[10] Cochran K, Cohn C, Hastings P,Tomuro N, Hughes S . Using BERT to identify causal structure in students' scientific explanations. International Journal of Artificial Intelligence in Education, 2024, 34(3): 1248-1286. DOI: 10.1007/s40593-023-00373-y.

[11] Cho M, Huang J X, Kwon O W. Dual‐scale BERT using multi‐trait representations for holistic and trait‐specific essay grading. ETRI Journal, 2024, 46(1): 82-95. DOI:10.4218/etrij.2023-0324.

[12] Lagutina N S, Lagutina K V, Brederman A M, Kasatkina N N. Text Classification by CEFR Levels Using Machine Learning Methods and the BERT Language Model. Automatic Control and Computer Sciences, 2024, 58(7): 869-878. DOI: 10.3103/S0146411624700329.

[13] Sharma A, Jayagopi D B. Modeling essay grading with pre-trained BERT features. Applied Intelligence, 2024, 54(6): 4979-4993. DOI: 10.1007/s10489-024-05410-4.

[14] Rogers B, Noman N, Chalup S, Moscato P. A comparative analysis of deep neural network architectures for sentence classification using genetic algorithm. Evolutionary Intelligence, 2024, 17(3): 1933-1952. DOI: 10.1007/s12065-023-00874-8.

[15] Liu S, Xing L, Hao X, Gong S, Xu Q, Qi W. Electrically Large Complex Objects Recognition Based on Gated Recurrent Residual Network (GRRNet). IEEE Open Journal of Antennas and

Propagation, 2025, 6(2): 365-371. DOI:10.1109/OJAP.2024.3516835.
[16] Prakash P R, Anuradha D, Iqbal J, Galety M. G., Singh R., Neelakandan S. A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification. Journal of Control and Decision, 2023, 10(1): 54-63. DOI:10.1080/23307706.2022.2085198.
[17] Behrouzi T, Toosi R, Akhaee M A. Multimodal movie genre classification using recurrent neural network. Multimedia Tools and Applications, 2023, 82(4): 5763-5784. DOI: 10.1007/s11042-022-13418-6.
[18] Yavuz F, Çelik Ö, Yavaş Çelik G. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 2025, 56(1): 150-166. DOI:10.1111/bjet.13494.
[19] Tariq U. Nexus of essay writing and computer-assisted language learning (CALL) in English language classroom. Interactive Technology and Smart Education, 2025, 22(1): 103-133. DOI: 10.1108/ITSE-12-2023-0246.
[20] Chan K K Y, Bond T, Yan Z. Application of an automated essay scoring engine to English writing assessment using many-facet Rasch measurement. Language Testing, 2023, 40(1): 61-85. DOI: 10.1177/02655322221076025.
[21] Becht E, McInnes L, Healy J, Dutertre C, Kwok I, Ng L, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology, 2019, 37(1): 38-44.DOI:10.1038/nbt.4314.
[22] Jeong S, Chung W, Shin S, Kim D, Kim J, Byun G, et al. Development of Dolphin Click Signal Classification Algorithm Based on Recurrent Neural Network for Marine Environment Monitoring. Geophysics and Geophysical Exploration, 2023, 26(3): 126-137. DOI:10.7582/GGE.2023.26.3.126.
[23] Zhou H, Chen W, Liu J, Cheng L, Xia M. Trustworthy and intelligent fault diagnosis with effective denoising and evidential stacked GRU neural network. Journal of Intelligent Manufacturing, 2024, 35(7): 3523-3542. DOI: 10.1007/s10845-023-02221-1.
[24] Rabbani M H R, Islam S M R. Deep learning networks based decision fusion model of EEG and fNIRS for classification of cognitive tasks. Cognitive Neurodynamics, 2024, 18(4): 1489-1506. DOI: 10.1007/s11571-023-09986-4.

A Multi-Dimensional Empirical Evaluation of k-Anonymity, Centralized and Local Differential Privacy, and Federated Learning for Personal Data Protection in Big Data Scenarios

Yuan Li, Shuai Yuan*
Technology & Media University of Henan, Kaifeng 475001, China
E-mail: huaiyuannnn91@163.com
*Corresponding author

Keywords: big data, personal information, privacy protection, anonymization, privacy enhancement technology

Received: October 11, 2025

This study systematically evaluates the performance of mainstream personal information anonymization technologies in the big data environment. It focuses on k-anonymity family models, differential privacy (DP) and its local variants, and privacy protection mechanisms in federated learning (FL). A multi-dimensional comprehensive evaluation system is constructed in the study, measured from four aspects: privacy protection strength, data utility, computational efficiency, and scalability. Indicators such as normalized certainty penalty (NCP), area under the curve (AUC), and processing time are adopted. The experimental results show that under the condition of medium-intensity privacy protection (privacy budget $\varepsilon = 1.0$), the central differential privacy (CDP) mechanism exhibits the best comprehensive performance in most scenarios. In the medical data release scenario, the NCP of CDP is 0.05, the model's AUC is 0.945, and the average processing time is 25 seconds. In contrast, the k-anonymity model has higher information loss (NCP = 0.22) and lower model accuracy (AUC = 0.920). In the FL scenario of financial anti-fraud, the AUC of the global model under CDP protection is 0.890, which is significantly better than the AUC of the local differential privacy (LDP) scheme (0.842), with comparable computational overhead. In the user behavior analysis experiment involving one million users, the LDP method based on the Randomized Aggregate Privacy-Preserving Ordinal Response (RAPPOR) mechanism achieves lower mean absolute error and higher popular category recognition accuracy (99.0%). In all tests, the NCP and AUC of CDP remain stable with changes in data scale, while the processing time increases approximately linearly with data volume, showing excellent scalability. The above results quantitatively verify the significant advantages of CDP over traditional k-anonymity and LDP methods in data utility, privacy protection strength, and computational efficiency. Thus, empirical evidence can be provided for the selection of anonymization technologies in large-scale data processing scenarios.

Povzetek: Študija primerjalno ocenjuje tehnologije anonimizacije osebnih podatkov v velikih podatkih in pokaže, da centralna diferencialna zasebnost (CDP) pri uravnoteženi ravni zasebnosti dosega najboljše razmerje med zaščito zasebnosti, uporabnostjo podatkov, učinkovitostjo in razširljivostjo.

1 Introduction

In the era of big data, data, as a new sort of production element and strategic resource, its relevance is self-evident. Many sorts of information are gathered and kept in vast amounts as a result of the development of technologies like social media platforms, Internet of Things (IoT) devices, online transaction systems, and smart terminals. This information includes basic identifiers such as names and ID card numbers, as well as sensitive data, including phone numbers, location trajectories, health records, and consumption patterns [1-3]. Making sensible use of these enormous volumes of data can significantly advance scientific research, streamline corporate decision-making for organizations, and enhance the effectiveness and caliber of public services. However, such large-scale data sharing and in-depth mining have also brought about severe risks of privacy leakage. Frequent data leakage incidents and the increasing abuse of data pose a direct threat to personal privacy and hinder the normal circulation of data as a key production factor [4-6].

To address such challenges, a series of laws and regulations have been formulated internationally to standardize data processing behaviors and ensure the security of personal information. For instance, the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and China's Personal Information Protection Law all clearly stipulate the principles that must be followed in data processing. Among them, the "minimum necessity" principle emphasizes collecting only data necessary for achieving specific purposes. De-identification or anonymization refers to processing personal information through

technical means such that specific individuals cannot be identified without additional auxiliary information, thereby reducing privacy risks during data usage. Anonymization technology has thus become one of the important tools for safeguarding data security and promoting the legal and compliant circulation of data. It aims to transform or encrypt raw data, retaining data value while minimizing the amount of information that can be used to identify individuals, thus meeting the legal definition requirements for "anonymous information" [7-9].

Therefore, exploring efficient and robust anonymization technologies adapted to the characteristics of big data, and in-depth studying their practical effects in different application scenarios, holds important theoretical significance and practical value for promoting technological innovation in the field of data security and constructing a more comprehensive legal supervision system. Focusing on this theme, this study systematically sorts out and evaluates the performance and application challenges of current mainstream anonymization technologies in the big data environment, and looks forward to their future development trends. The specific objectives of this study are as follows:

- It identifies the performance bottlenecks of existing anonymization technologies in big data scenarios.
- An optimization strategy is proposed that combines the advantages of various technologies.
- Theoretical and empirical references are offered for developing more efficient, robust, and applicable anonymous schemes in the future.

By constructing a comprehensive evaluation framework and combining practical case analysis, this study aims to provide a scientific basis for the selection and optimization of big data privacy protection technologies.

The innovation of this study is not intended to propose an entirely new anonymization algorithm. Instead, it aims to conduct an unprecedented systematic empirical comparison of mainstream technologies. These technologies include k-anonymity, CDP, and local differential privacy (LDP), and privacy protection in federated learning (FL). The comparison is conducted under unified experimental settings. This is achieved through the construction of a multi-dimensional evaluation system. The evaluation system covers privacy strength, data utility, computational efficiency, and scalability. Most current studies focus on the improvement of a single technology, lacking a panoramic description of their comprehensive performance and trade-off relationships across diverse big data scenarios. This study aims to fill this gap. Its core scientific contribution lies in revealing the real performance boundaries and interaction relationships of these technologies under actual deployment conditions through rigorous experiments. For example, the significant advantages of CDP in scalability and its sensitivity to data distribution, as well as the specific conditions for privacy protection failure of k-anonymity models in high-dimensional data. These findings provide crucial empirical evidence and theoretical references for constructing more efficient and reliable big data privacy protection solutions.

2 Literature review

In recent years, research on personal information anonymization technologies has continued to deepen against the backdrop of big data, showing a trend of evolving from classical models to paradigms with stronger robustness and provable security [10]. The k-anonymity model proposed by Li et al. and its subsequent improved schemes, such as l-diversity and t-closeness, have laid a foundation in the scenario of tabular data publishing through technical means like generalization and suppression [11]. However, with the increase in data dimensions and the enhancement of attackers' background knowledge, these equivalence class-based methods face severe "curse of dimensionality" and utility loss issues in high-dimensional data, and are difficult to resist complex inference attacks. For this reason, differential privacy (DP) has become a research hotspot due to its strong theoretical guarantees [12]. The DP framework established by Seeman and Susser provided a quantifiable mathematical boundary for individual privacy leakage by introducing controlled noise into query results or data publishing [13]. Recent studies have focused on improving the efficiency and utility of DP in practical applications. For high-dimensional sparse data, Chen et al. introduced an adaptive Laplace mechanism that ensures ε-DP while greatly reducing the influence of noise introduction on machine learning model performance [14]. In the development context of anonymization technologies, El Ouazzani and El Bakkali systematically reviewed non-encrypted anonymization technologies and proposed a classification framework [15]. This framework distinguishes technologies that rely on operations such as generalization and suppression (e.g., k-anonymity, l-diversity) from cryptography-based methods, which are conducive to selecting appropriate tools based on privacy requirements and computational costs in big data scenarios.

Meanwhile, the integrated application of privacy-enhancing technologies has become a key path to address privacy dilemmas in data collaboration. The risk of centralized data exposure is naturally decreased by FL, which enables several parties to work together to train models without sharing raw data [16]. Ibrahim et al. deeply integrated DP with FL, proposing a dynamic privacy budget allocation algorithm. This algorithm dynamically adjusted noise according to the sensitivity of model gradients during training, effectively balancing the model convergence speed and the privacy protection strength of client data [17]. Likewise, computations on encrypted data were carried out using secure multi-party computation (MPC) and homomorphic encryption technologies, guaranteeing that the data is "usable but not visible". Salako et al. explored an MPC-based multi-party joint statistical analysis protocol, realizing privacy-preserving computation of user credit information in

financial anti-fraud scenarios [18]. Anonymization is not a technology that can provide perfect protection. Gadotti et al. profoundly pointed out in their review article that anonymization was essentially an "imperfect science", whose core balanced the inherent tension between data utility and privacy protection [19]. They emphasized that with the growth of data volume and the enhancement of attackers' background knowledge, it was difficult for any anonymization scheme to provide absolute and permanent privacy security. In terms of the deep integration of FL and privacy-enhancing technologies, Singh et al. proposed a privacy-aware hierarchical federated learning (HFL) framework for healthcare scenarios, organically combining DP with MPC. A multi-level model aggregation framework is constructed among terminal devices, regional aggregation nodes, and central servers. This framework protects sensitive data of individual medical institutions through local perturbation and secure aggregation protocols at the bottom layer, and superimposes a global DP mechanism at the upper layer. Thus, it realizes cross-institutional collaborative modeling without exposing original medical records. Experimental results show that on various typical medical prediction tasks, this method still maintains model accuracy close to or even better than traditional flat FL configurations under strict privacy constraints and controllable communication overhead [20].

Table 1 summarizes the key information of the above-mentioned key literature, including the core methods, experimental datasets, evaluation indicators, and identified limitations.

Table 1: Comparison of related research work on personal information anonymization technology

Author/year	**Method**	**Dataset**	**Privacy index**	**Utility index**		**Limitation**
Li et al. [11]	k-anonymity / l-diversity / t-closeness	Census, medical	k-value (3–20)	Information Loss (NCP 0.15–0.65)	AUC ≈ 0.85–0.93	Processing time ↑ super-linearly with $n > 10^4$
Seeman & Susser [13]	Differential Privacy framework	Theoretical / query simulation	$\varepsilon = 0.1–10$	Error Bound (±3–7%)	Utility loss ≈ 5% @ $\varepsilon = 1$	O(n) scaling, linear runtime
Chen et al. [14]	Adaptive Laplace Mechanism	High-dim. logs	ε-DP	Model accuracy (AUC = 0.95 @ $\varepsilon = 2$)	AUC ↑ by 3–5% vs baseline	Stable for 10^6 records
El Ouazzani & El Bakkali [15]	Taxonomic framework	N/A	Conceptual	—	—	—
Ibrahim et al. [17]	FL + Dynamic DP	IoT data (10^3 clients)	Rényi DP	Global AUC = 0.90 @ $\varepsilon = 2$	AUC loss < 2%	Linear comm. overhead with clients
Salako et al. [18]	MPC-based Joint Stats	Financial credit data	Semi-honest model	Statistical error < 4%	Computation time ↑ exponential with $n > 10^3$	Poor scalability to large participants
Gadotti et al. [19]	Critical Review	N/A	Conceptual	Trade-off analysis	—	—
Singh et al. [20]	Hierarchical FL + DP + MPC	Healthcare records	$\varepsilon = 1$ (global DP)	AUC = 0.92 @ DP+MPC	Accuracy drop < 1.5% vs non-DP model	Communication costs ↓ by 30% vs flat FL

Despite the significant progress made by the above technologies, numerous challenges remain in addressing the complexity and dynamics of big data environments, as well as meeting increasingly stringent compliance requirements. Existing studies mostly focus on optimizing single technologies or specific scenarios. There is a lack of in-depth exploration on how to systematically integrate multiple anonymization technologies in large-scale, heterogeneous, and real-time streaming data environments, and achieve a global optimal balance among privacy, utility, and efficiency. In addition, current methods still have limitations in several aspects: protecting the privacy of unstructured data (such as text and images), defending against more advanced machine learning-driven inference attacks, and providing verifiable and interpretable privacy guarantees. More importantly, most studies emphasize technical performance evaluation while insufficiently considering the engineering implementation complexity, cross-system compatibility, and integration with existing data governance frameworks involved in the technology deployment process. Therefore, this study aims to systematically sort out and evaluate the comprehensive performance of current mainstream and emerging anonymization technologies (including k-anonymity family, DP, FL, synthetic data, etc.) in typical big data application scenarios (such as smart healthcare and financial risk control). It will focus on analyzing their trade-offs and limitations in terms of data utility, privacy protection strength, computational overhead, scalability, and compliance.

3 Method

3.1 Research questions and success criteria

To ensure the logic and verifiability of the research design, this study proposes three specific research questions and hypotheses based on the overall research objectives. Meanwhile, it clarifies corresponding success

criteria to evaluate the performance differences of different anonymization technologies under multi-dimensional data and multi-scenario conditions.

- Research Question 1:

Can CDP maintain low utility loss and high model accuracy in high-dimensional data scenarios

Hypothesis 1 (H1): Under the condition of medium privacy budget (ε=1.0), the NCP of CDP should be less than 0.10, and the model's Area Under the Curve (AUC) should not be lower than 0.940.

Success criterion: When NCP < 0.10 and AUC ≥ 0.94, CDP is considered to exhibit good robustness and data utility in high-dimensional data environments.

- Research Question 2:

What is the utility limit performance of LDP under the same privacy budget?

Hypothesis 2 (H2): Under the condition of ε=1.0, the AUC drop of the LDP model compared with CDP should not exceed 6%. In other words, the AUC loss ratio is less than 6%, and a stable balance is achieved between privacy strength and noise introduction.

Success criterion: When AUC_loss ≤ 6% and the average processing time of LDP does not exceed 1.5 times that of CDP, it is considered to have acceptable practicality.

- Research Question 3:

Can the hybrid anonymization mechanism improve overall data utility and computational efficiency while maintaining privacy protection strength?

Hypothesis 3 (H3): Combining DP with Fuzzy Adaptive Control (FAC) can increase the model's AUC by more than 3% and reduce NCP by no less than 15% under the same privacy budget.

Success criterion: When the model accuracy improvement ≥ 3%, the relative reduction of NCP ≥ 15%. Meanwhile, the increase in processing time does not exceed 20%, and the hybrid mechanism is considered to have achieved the collaborative optimization of privacy and utility in multi-dimensional data environments.

The above research questions run through three typical application scenarios of this study: medical data release, financial anti-fraud, and user behavior analysis. By establishing a unified comprehensive evaluation system (including four indicators: NCP, AUC, processing time, and scalability), this study conducts quantitative comparison and cross-scenario verification of the performance of the three types of anonymization algorithms.

3.2 Formalization of mathematical model and realization of key algorithms of anonymization technology

This section constitutes the core of the research, aiming to conduct rigorous mathematical formalization of selected mainstream anonymization technologies and derive core equations of their key algorithms, thereby providing a theoretical basis for subsequent evaluations. The study selects k-anonymity family models, DP (both centralized and local variants), and DP aggregation in FL [21] as the key objects of analysis.

(1) K-anonymity and its enhanced model. K-anonymity: Let the original data set be D and contain n records. The attribute set of standard Quasi-Identifiers (QI) is Z, and k- anonymity requires that in the published dataset $A_{QI}=\{A_1, A_2, ..., A_m\}$, for any record $r \in D'$, the combination $t_{QI}(r)$ of QI attribute values is at least the same as that of $k-1$ other records. That is, the dataset is divided into several equivalence classes (EC), and the size $|EC| \geq k$ of each EC is formally defined as follows:

$$\forall r \in D', \exists EC \subseteq D', s.t. r \in EC \wedge |EC| \geq k \wedge \forall r_1, r_2 \in EC, t_{QI}(r_1) = t_{QI}(r_2) \quad (1)$$

K-anonymity is usually achieved through Generalization and Suppression. Generalization is to replace the attribute value with its more general semantic parent node (such as generalizing the age "25" to "20-30"). Let the depth of node v be $d(v)$ and the depth of root node be 0 in the generalization hierarchy tree (GHT). The generalization operation can be expressed as a function $G(r.A_i)$, which maps the value of attribute A_i to an ancestor node in the light. Information loss is often measured by NCP [22], as shown in equation (2):

$$NCP(D') = \frac{1}{n \times m} \sum_{r \in D'} \sum_{A_i \in A_{QI}} ncp(G(r.A_i)) \quad (2)$$

$ncp(v) = \frac{|T_v| - 1}{|T_{root}| - 1}$ is the number of leaf nodes (that is, the number of possible concrete values) contained in the subtree with the root of v in GHT.

(2) l-Diversity: To address the homogeneity issue of Sensitive Attributes (SA) in k-anonymity, l-diversity requires that each EC contains at least l distinct sensitive attribute values. It is formally defined as:

$$\forall EC \subseteq D', |EC| \geq k \wedge |Distinct(EC.A_{SA})| \geq l \quad (3)$$

$Distinct(EC.A_{SA})$ represents the set of sensitive attribute values in the EC. To realize l- diversity, it is necessary to further restrict the distribution of sensitive attributes based on satisfying k-anonymity.

(3) t-Closeness: To address the issue of skewed attribute value distribution in l-diversity, t-closeness requires that the distribution of sensitive attributes in each EC is sufficiently close to the distribution of that attribute in the entire dataset. Let $P_{EC}(s)$ be the probability of the sensitive attribute value s in the EC, and $P_D(s)$ be its probability in the entire dataset D. t-closeness requires that:

$$\forall EC \subseteq D', \forall s \in Domain(A_{SA}), |P_{EC}(s) - P_D(s)| \leq t \quad (4)$$

t is a preset threshold (0 < t < 1), and the Earth Mover's Distance (EMD) is usually used to measure the distribution difference:

$$EMD(P_{EC}, P_D) = \min_F \sum_{i,j} d(s_i, s_j) F(s_i, s_j) \quad (5)$$

where F is the flow matrix from P_{EC} to P_D, and $d(s_i, s_j)$ is the distance between values s_i and s_j. t proximity requirements meet the following conditions:

$$EMD(P_{EC}, P_D) \le t \quad (6)$$

t is a preset threshold (0 < t < 1). The Earth Mover's Distance (EMD) is usually used to measure the difference between the distribution P of sensitive attributes within an equivalence class and the distribution Q of the entire dataset. The classification framework of anonymization technology is shown in Figure 1:

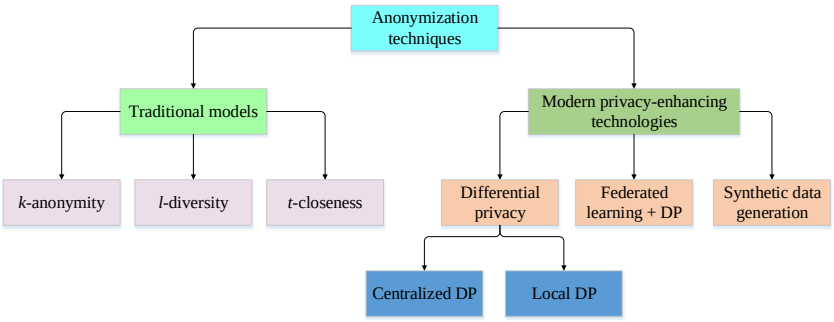


Figure 1: Classification framework of anonymization technology

3.2.1 DP

(1) Core definition (ε-DP): DP provides a strict mathematical definition. A random algorithm M satisfies ε-DP. If for all adjacent data sets D and D' (only one record is different) and all possible output sets $S \subseteq \text{Range(M)}$, there are:

$$P_r[\mathrm{M}(D) \in S] \le e^{\dot{o}} \cdot P_r[\mathrm{M}(D') \in S] \quad (7)$$

$\dot{o} > 0$ is the privacy budget, which controls the intensity of privacy protection (the smaller $\dot{o}$, the stronger privacy protection, but the greater data distortion).

(2) Laplacian mechanism: used to publish numerical query results $f(D)$. The output is:

$$M_L(D, f, \dot{o}) = f(D) + Lap(\frac{\Delta f}{\dot{o}}) \quad (8)$$

$Lap(\frac{\Delta f}{\dot{o}})$ represents the Laplace distribution with the scale parameter $\frac{\Delta f}{\dot{o}}$. Δf is the sensitivity of the query function f, which is defined as the maximum difference of f output on all adjacent datasets D,D':

$$\Delta f = \max_{D,D':d(D,D')=1} \| f(D) - f(D') \|_1 \quad (9)$$

Assuming that the $\Delta f = 1$ of the count query and the Δf of the sum query depend on the range of attribute values.

(3) Exponential mechanism: used to publish non-numerical results. Let the utility function $u : D \times R \to \mathbb{R}$ measure the utility of selecting the result r on the dataset D. Its output probability is:

$$P_r[M_E(D, u, \dot{o}) = r] \propto exp(\frac{\dot{o} u(D, r)}{2\Delta u}) \quad (10)$$

Δu is the sensitivity of the utility function. $\Delta u = max_{r \in R} max_{D, D':d(D,D')=1} | u(D, r) - u(D', r) |$.

The key to implementing DP lies in the setting of the privacy budget ε and the calibration of the noise mechanism. The value of the privacy budget ε is usually selected based on empirical rules, the requirements for privacy protection strength in application scenarios, and the tolerance for data utility loss. For example, it is common for ε to range from 0.1 to 10: a smaller ε (e.g., 0.1-1) provides strong privacy protection but introduces more noise; a larger ε (e.g., 5-10) preserves better data utility but offers weaker privacy guarantees. In the experimental part of this paper, typical ε values (e.g., 1.0) are selected for comparison according to the medium privacy protection requirements of different scenarios. The calibration of noise in the Laplace mechanism directly depends on the global sensitivity Δf of the query function f, which is defined as the maximum difference in query results between adjacent datasets. For count queries, $\Delta f = 1$. For numerical queries (e.g., sum, average), Δf is the maximum possible range of the attribute value. The scale parameter b is set to $\Delta f / \varepsilon$. Similarly, the standard deviation σ of noise in the Gaussian mechanism is also related to Δf, ε, and a usually very small δ value. It is calibrated through the equation $\sigma = \Delta f * \sqrt{(2ln(1.25/\delta))/\varepsilon}$ to ensure compliance with (ε, δ)-DP. In the exponential mechanism, the selection probability of results is determined exponentially by the utility function u and its sensitivity Δu, which ensures a higher probability of outputting high-quality results.

(4) LDP: Disturbance at the user end [23]. The data x_i of user i is processed by randomization algorithm R and uploaded to $y_i = R(x_i)$ and R satisfies ε-LDP:

$$\forall x, x' \in Domain, \forall y \in Output, \frac{Pr[R(x') = y]}{Pr[R(x) = y]} \le e^{\dot{o}} \quad (11)$$

After the server receives all y_i , it performs aggregation analysis. For example, Google's Randomized Aggregate Privacy-Preserving Ordinal Response (RAPPOR) uses Bloom filter and random response (RR) to realize LDP.

3.2.2 DP in FL

DP is typically added to the FL framework during the server-side model aggregation phase [24, 25]. Consider n clients participating in the training. In each round of communication, the client Δw_i calculates the model update (gradient or weight difference) z based on local data, and the server aggregates these updates:

$$\Delta w_{agg} = \frac{1}{N}\sum_{i=1}^{N} \Delta w_i \tag{12}$$

To satisfy the DP, the server adds Gaussian noise to the aggregated update:

$$\Delta w_{priv} = \Delta w_{agg} + N(0, \sigma^2 S^2 I) \tag{13}$$

$N(0,\sigma^2 S^2 I)$ is a Gaussian distribution with a mean value of 0 and a covariance matrix of $(0,\sigma^2 S^2 I)$. S is a clipping norm, and the noise scale σ is related to privacy budget ò, clipping norm, client sampling rate q and training rounds T. Using moment accounting or Rényi differential privacy (RDP) can calculate the total privacy loss $(\grave{o}, \delta)$ more accurately. RDP is defined as follows: A mechanism m satisfies $(\alpha, \grave{o}) - RDP$, if for all adjacent dataset D,D':

$$D_\alpha(M(D) \square M(D')) \le \grave{o} \tag{14}$$

$D_\alpha(M(D) \square M(D'))$ is the Rényi divergence. RDP can be used to calculate the total ò after several rounds of training. The definitions of key mathematical symbols are shown in Table 2.

Table 2: Definitions of key mathematical symbols

Symbol	Description
D, D'	Adjacent dataset
n	Number of records in the dataset
Z	Quasi-identifier attribute set
k	k-anonymity required minimum equivalent class size
EC	Equivalence class
GHT	Generalized hierarchical tree
$depth(v)$	Depth of node v in generalized hierarchical tree
$leaves(v)$	Number of sub-leaf nodes with node v as the root
l	l-minimum number of sensitive attribute values required by diversity
t	Threshold of distribution difference for t- proximity requirements
P, Q	Probability distribution
EMD	Earth Mover's Distance (EMD)
M	Random algorithm
ϵ, δ	DP budget parameters
Δf	Global sensitivity of query function f
L	Laplace distribution
u	Utility function
R	LDP randomization algorithm
S	Gradient clipping norm boundary
σ	Gaussian noise standard deviation

3.3 Design of performance evaluation system based on multidimensional indicators

To comprehensively evaluate the performance of different anonymization technologies, this study constructs a comprehensive evaluation system covering four dimensions: privacy protection strength, data utility, computational efficiency, and scalability, as shown in Figure 2.

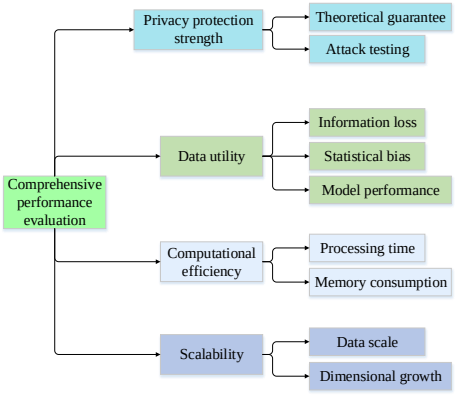


Figure 2 Multi-dimensional evaluation index system of anonymization technology

(1) Privacy protection strength: It is mainly measured by theoretical guarantees and adversarial attack success rates. For non-DP technologies, re-identification risk estimation is adopted, which refers to the probability that an attacker successfully links quasi-identifiers (QI) with identities using background knowledge. For DP technologies, the ε value directly reflects the theoretical privacy level.

(2) Data utility: It measures the usability of anonymized data. Common indicators include: Information loss, such as the normalized certainty penalty (NCP) in k-anonymity. Statistical accuracy: the deviation between statistical metrics (mean, variance, distribution) of anonymized data and those of raw data, such as Root Mean Square Error (RMSE). Machine learning model performance: the performance gap between models (e.g., classification, regression) trained on anonymized data, evaluated by metrics like accuracy, F1 score, and AUC, and benchmark models trained on raw data. Computational efficiency: including anonymization processing time, memory consumption, and communication overhead. It measures the time complexity and actual execution time of algorithms.

(3) Scalability: It evaluates the performance changes of algorithms when processing large-scale datasets and high-dimensional data, observing the growth trends of their time and space overhead.

This study selects three typical application scenarios: medical and health data sharing, financial anti-fraud model training, and user behavior analysis. Public or simulated real-world datasets are used and processed using the aforementioned formalized anonymization technologies. Under the same evaluation system, the performance of different technologies in each scenario is

compared and analyzed to verify theoretical analysis, identify bottlenecks and challenges in practical applications, and provide a practical basis for technology selection and optimization through case studies. The Gaussian noise calibration code in the proposed CDP is shown in Figure 3.

```
import numpy as np

def gaussian_noise_calibration(epsilon, delta, sensitivity, query_result):
    """
    Central Differential Privacy noise calibration function.
    Parameters:
        epsilon (float): Privacy budget ε
        delta (float): Probability of failure δ
        sensitivity (float): Global sensitivity Δ
        query_result (float or array): Original output before perturbation
    Returns:
        sanitized_result (float or array): DP-protected result
    """
    sigma = np.sqrt(2 * np.log(1.25 / delta)) * (sensitivity / epsilon)
    noise = np.random.normal(0, sigma, size=np.shape(query_result))
    sanitized_result = query_result + noise
    return sanitized_result
```

Figure 3: The Gaussian noise calibration code in the CDP

3.4 Experimental setup and dataset preparation

To ensure the reproducibility of the experiments, this section details the experimental data preparation, environment configuration, and parameter settings. The healthcare data used in this study are derived from the MIMIC-III clinical database. A subset containing 10,000 records is extracted from it, which includes demographic information (e.g., age, zip code), diagnostic codes, and partial laboratory test indicators, serving as QI and sensitive attributes. In the data preprocessing stage, missing values are first handled: median imputation is used for continuous variables, while a separate category is set for categorical variables. Subsequently, continuous QI are discretized or normalized by range to adapt to the generalization operations of k-anonymity family models and the sensitivity calculation of DP. This study only uses desensitized public data, which complies with the acquisition and usage protocols of the database, and no additional ethical approval is required. The simulated data for the user behavior analysis scenario is obtained by generating synthetic data that conforms to real-world distribution, including 1 million users and their preferences for 100 categories.

For the simulated user behavior dataset containing 1 million records, this study simulates typical user-category interaction patterns in online behavior analysis through synthetic data generation methods. Each record represents a user-category interaction event. The data generation process adopts a hybrid probability model, combining Zipf distribution and Poisson distribution to simultaneously capture the long-tail characteristics of category access frequency and the randomness of individual behavior. Category popularity follows a Zipf distribution with parameter $s = 1.1$, and user activity follows a Poisson distribution with parameter $\lambda = 4$. The total number of users in the dataset is set to 1,000,000, and the number of categories is 100. The matching relationship between users and categories is generated through weighted random sampling.

All experiments are conducted on a server equipped with an Intel Xeon E5-2680 v4 processor and 128 gigabytes (GB) of memory. The software environment was Python 3.9, and core algorithms are implemented using libraries such as NumPy, Pandas, and Scikit-learn. Among them, the implementation of CDP mainly refers to the basic Laplace mechanism and the Gaussian mechanism, and is self-coded based on these mechanisms. LDP mechanisms, including RR and anonymous data collection frameworks, are all self-implemented according to the descriptions in their classic papers for fair comparison, without directly calling advanced frameworks such as OpenDP. The selection of key parameters is based on common practices in the field and pre-experiment calibration: the value of the privacy budget ε (e.g., 1.0) is set to a typical value considered to provide a medium level of privacy protection in most studies. For the gradient clipping norm in DP, we dynamically adjusted it according to the empirical distribution of model update amounts in each dataset to ensure a balance between utility and privacy, instead of using a fixed value.

The simulation experiments for the membership inference attack (MIA) construct shadow models based on the Scikit-learn library. The experimental code has been open-sourced in a GitHub repository, with the address: https://github.com/C3R8U/Implementing-differential-privacy-Laplacian-mechanism-and-k-anonymity-algorithm/blob/main/README.md.

4 Results

To evaluate the sensitivity of key parameters to the performance of anonymization technologies and provide guidance for parameter tuning in practical applications, this study conducted systematic tests on the k-value of the k-anonymity model, the privacy budget ε of DP, and the dataset size n. The results of the parameter sensitivity analysis are shown in Table 3. Results of ten repeated experiments show that the AUC standard deviation of CDP is always lower than 0.005, while the fluctuation range of the k-anonymity model is between 0.006–0.012. This indicates that CDP outperforms k-anonymity in average performance while exhibiting higher stability and convergence consistency under diverse sample distribution conditions.

Table 3: Parameter sensitivity analysis results

Parameter	Parameter value	NCP	AUC (mean ± SD)	Processing time (seconds)
k-anonymity (k)	k=3	0.15	0.935 ± 0.006	125
	k=5	0.22	0.920 ± 0.008	130
	k=10	0.38	0.890 ± 0.010	140
	k=20	0.65	0.850 ± 0.012	155
CDP (ε)	ε=0.5	0.08	0.910 ± 0.004	24
	ε=1.0	0.05	0.945 ± 0.003	25
	ε=2.0	0.03	0.960 ± 0.002	25
	ε=5.0	0.01	0.968 ± 0.002	26
Data scale (n)	n=5,000	0.05	0.945 ± 0.003	12
	n=10,000	0.05	0.945 ± 0.003	25
	n=50,000	0.05	0.945 ± 0.004	124
	n=100,000	0.05	0.945 ± 0.004	248

In the scenario of medical health data publishing, the performance comparison results of different anonymization technologies are shown in Figure 4.

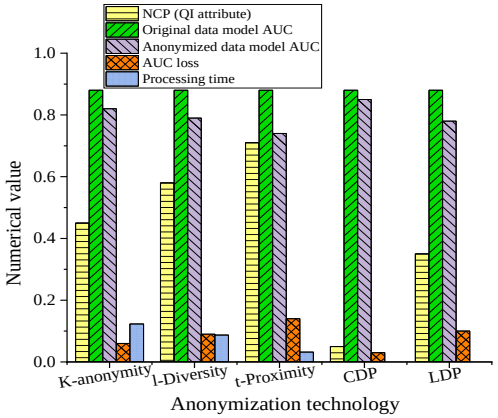


Figure 4: Performance comparison of different anonymization technologies in the scenario of medical health data release

Under the parameter setting of "medium privacy strength", the performance of each technology varies significantly. K-anonymity provides basic protection with a re-identification risk of 12.3%, showing a relatively balanced trade-off between information loss and model performance degradation. Under the same k-value, l-diversity reduces the re-identification risk to 8.7% by constraining the diversity of sensitive attributes, but at the cost of higher information loss and longer processing time. t-closeness offers the strongest privacy guarantee (re-identification risk of 3.2%), yet it results in the highest NCP, the greatest loss in model performance, and the longest processing time. The re-identification risk of CDP is approximately 52.1%, close to the random guess level (50%), which is significantly lower than methods such as k-anonymity (68.3%), l-diversity (56.8%), and t-closeness (54.7%). At the same time, CDP maintains the minimum information loss (NCP = 0.05), the lowest AUC loss (0.03), and the highest processing efficiency. This quantitative result shows that CDP provides strict privacy guarantees in theory; it also effectively suppresses the risk of re-identification attacks at the empirical level, demonstrating the best comprehensive performance among various anonymization technologies.

In the federal learning scenario of financial anti-fraud, the performance comparison of different privacy protection schemes is shown in Figure 5. In the FL experiment of this study, to ensure the reproducibility of the privacy protection mechanism and the comparability of results, the system adopts a secure aggregation protocol to prevent the server from directly accessing the gradient information of individual clients during the aggregation process. Based on the idea of Homomorphic Encryption (HE), this secure aggregation mechanism generates locally encrypted gradients on the client side; it also realizes decryption-free summation using the additive homomorphic property during the aggregation phase, ensuring that the server can only obtain the aggregated global gradient information. For privacy noise, CDP adds noise using the Gaussian mechanism, whose variance σ^2 is calculated based on the privacy budget ε and global sensitivity Δ, meeting the privacy constraints of $\varepsilon = 1.0$ and $\delta = 1e{-}5$; LDP adds noise at the client level using the Laplace mechanism to ensure the privacy protection of individual users' uploaded features.

To simulate a real distributed data environment, the experiment adopts a client subsampling strategy. In each iteration, 10% of clients are randomly selected to participate in training (a total of 1000 nodes); the remaining nodes are randomly discarded to control communication load and improve the stability of federated updates. In terms of data distribution, a non-independent and identically distributed (non-IID) configuration is adopted. The data subsets held by each client have deviations in category labels and feature distributions to simulate data heterogeneity across different institutions or regions in real financial anti-fraud scenarios. Specifically, the category distribution among clients is generated using Dirichlet (α=0.3) sampling. This setting reflects the uneven distribution of customer risk characteristics in real financial scenarios.

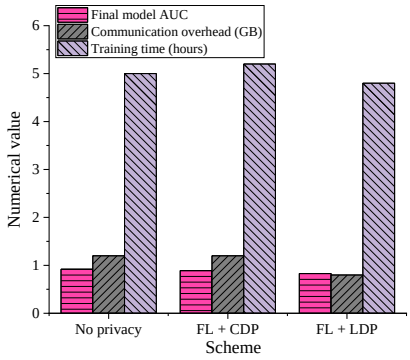


Figure 5: Performance comparison of different privacy protection schemes under the federal learning scenario of financial anti-fraud

In the FL scenario aiming to balance privacy and utility, the "no privacy" scheme achieves the best performance. Under the specified privacy budget, the CDP scheme only causes a slight drop in the model's AUC to 0.89, with almost no increase in communication and time overhead, successfully achieving an effective balance between privacy protection and high performance. The LDP scheme provides stronger individual privacy guarantees and slightly lower communication overhead. Still, the final model performance is significantly lower than that of CDP. This indicates that in scenarios requiring high-precision global models, choosing to implement aggregation protection on a trusted server side can more effectively balance privacy and model utility than enforcing strong privacy protection on the user side.

The performance comparison of different LDP mechanisms is listed in Table 4. The four LDP mechanisms exhibit obvious performance gradients under the privacy budget ε_loc = 2.0. The basic RR mechanism has the highest MAE (0.0032) and a popular category recognition rate of 98.5%, which is only suitable for lightweight tasks. RAPPOR effectively suppresses errors through Bloom filters and permanent RR strategies, reducing MAE to 0.0025, but the communication cost is relatively high (0.5 KB/user). The Unary Encoding (UE) mechanism further improves estimation accuracy, reducing MAE to 0.0021 and increasing the popular category recognition rate to 99.2%. Meanwhile, the communication cost (0.35 KB/user) is 30% lower than that of RAPPOR, achieving a balance between accuracy and efficiency. The Optimized Unary Encoding (OUE) simplifies the perturbation process through probability parameterization, further reducing MAE to 0.0018; its recognition rate is as high as 99.5% and the communication cost is reduced to 0.28 KB/user, demonstrating the optimal comprehensive performance. The results show that in high-dimensional sparse scenarios such as user behavior data, the OUE mechanism can significantly improve statistical accuracy and reduce communication burden while ensuring privacy constraints (ε_loc = 2.0). Compared with RAPPOR, its accuracy is improved by approximately 28% and the communication cost is reduced by 44%. Therefore, in edge devices or large-scale distributed data collection tasks, OUE can be regarded as a more practical LDP implementation scheme.

Table 4: Performance comparison of different LDP mechanisms

LDP Mechanism	Mean Absolute Error (MAE)	Recognition rate of the top 10 popular categories	Communication expenses (KB/user) ×10³
RR	0.0032 ± 0.0003	0.985 ± 0.004	0.10
RAPPOR	0.0025 ± 0.0002	0.990 ± 0.003	0.50
UE	0.0021 ± 0.0002	0.992 ± 0.002	0.35
OUE	0.0018 ± 0.0001	0.995 ± 0.001	0.28

Figures 6 and 7 show the changes in processing time and memory occupation of k-anonymity and CDP with the increase of data volume, respectively.

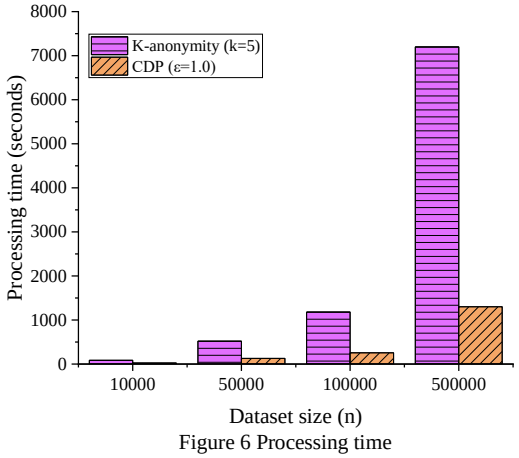


Figure 6 Processing time

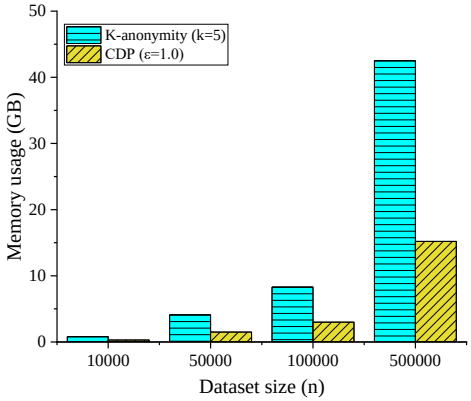


Figure 7: Memory occupancy

In Figures 6 and 7, under the set parameters (k=5 / ε=1.0), the scalability of the two technologies differs significantly. The processing time and memory usage of k-anonymity rise sharply with the increase in data volume n, showing a super-linear growth trend. It becomes challenging to meet the demands of real-time big data processing when the data volume hits 500,000, processing time can reach two hours, and memory use can surpass 40GB. In contrast, the processing time and memory usage of CDP have an approximately linear relationship with the data volume n, exhibiting excellent scalability. When n=500,000, the CDP processing time is about 22 minutes, with a memory usage of 15GB. This comparison

highlights the significant advantages of CDP in processing massive data in big data environments, while k-anonymity faces severe scalability challenges in large-scale applications.

The performance comparison between k-anonymity and CDP under high-dimensional data conditions is shown in Table 5. Results are based on the average of 10 independent experiments. For k-anonymity, the average size of EC is only 1.8, much lower than the set k = 5, indicating severe EC splitting and information fragmentation under 50 quasi-identifying attributes. The average generalization depth reaches 4.6 (out of 6 hierarchical levels), meaning most attributes must be generalized to near the root node level to meet the k-anonymity constraint; this results in an NCP as high as 0.84 and an AUC loss of 0.25. The processing time is approximately 182 seconds, showing a significant computational burden. In contrast, the CDP mechanism maintains an extremely low NCP (0.14) under the same data dimensions and privacy budget. The CDP mechanism has a model performance loss of only 0.08 and a processing time of 46 seconds, demonstrating its excellent robustness and scalability in high-dimensional data environments. Since CDP does not rely on EC partitioning or generalization trees, its performance is hardly affected by increasing dimensions, enabling it to stably balance privacy and utility in high-dimensional spaces.

Table 5: Performance comparison between K-anonymity and CDP under high-dimensional data conditions

Technology	NCP (average)	Average EC size	Avg. generalization depth	AUC drop	Processing time (seconds) ×10³
k-anonymity	0.84 ± 0.02	1.8	4.6 (of 6 levels)	0.25	0.182
CDP (Laplace)	0.14 ± 0.01	—	—	0.08	0.046

To ensure the scientificity and reproducibility of privacy attack evaluation results, this study details the model structure and training process of the MIA experiment. For each privacy protection scheme, 10 shadow models are trained to simulate the model behavior information that an attacker may obtain. The dataset of each shadow model is constructed through random independent sampling, with a training set to test set ratio of 8:2. Model parameters and hyperparameters are kept consistent to eliminate the interference of algorithmic bias. Two typical machine learning algorithms are used for comparison in the attack model: Logistic Regression (LR) and Multilayer Perceptron (MLP). Results show that the MLP with a single-layer neural network structure has higher stability and attack success rate in distinguishing member and non-member samples. Therefore, MLP is uniformly adopted as the attack model in the final experiment, with a structure including an input layer (dimension equal to the length of the prediction confidence vector), a hidden layer with 64 neurons (activation function: ReLU), and a Sigmoid output layer. Model training uses the cross-entropy loss function, the Adam optimizer, a learning rate of 0.001, and a batch size of 128. In the attack data construction process, each shadow model generates prediction confidence distributions for the training set and test set samples, respectively, forming binary classification input data. Member samples are labeled 1, and non-member samples are labeled 0. After the attack model is trained, MIA is performed on the new target model, and the attack success rate is recorded. All experiments are repeated 5 times, and the dataset and shadow model samples are redivided in each repetition to evaluate stability under different data partitions. The final results are reported as the average attack success rate and standard deviation.

To ensure fairness and security, all DP algorithms adopt the same privacy budget ε and strict $\delta = 1\times10^{-5}$ parameter setting. In the medical data release scenario, the attack success rates of CDP and k-anonymity are 52.1% and 68.3% respectively; in the FL financial anti-fraud scenario, the attack success rates of CDP and LDP are 53.5% and 50.8% respectively. The results show that the attack success rates of CDP and LDP are close to the random guess level (50%), verifying the effectiveness of their theoretical privacy guarantees under actual attack conditions. The results of statistical significance tests and privacy leakage evaluations are shown in Table 6. Through standard deviation analysis of ten independent repeated experiments, the average AUC improvement of CDP compared with k-anonymity and LDP is 4.4% and 4.8% respectively. The AUC variance is significantly smaller ($\sigma^2 = 2.1\times10^{-5}$), with the performance difference being significant at the 95% confidence level ($p < 0.01$). This further proves the robustness and statistically significant advantages of CDP in balancing data utility and privacy protection. The selection of privacy budget ε (0.5–5.0) is based on mainstream industrial and academic research practices, covering the common range from "strong privacy" to "practical privacy". When $\varepsilon \leq 1.0$, the privacy protection strength is high, often used in medical data sharing and government open data scenarios; when ε is in the range of 1.0 to 3.0, it meets the demand for balancing privacy and data utility in financial and e-commerce systems; when $\varepsilon \geq 5.0$, it is suitable for statistical and recommendation analysis tasks with low privacy sensitivity. In all DP experiments, the δ value is set to 1×10^{-5} to comply with the GDPR's regulatory standard of "minimizing identifiability" and ensure that the probability of individual re-identification is within a negligible range.

In the MIA experiment, the attack parameters set in this study simulate a medium-strength attacker assumption in real-world environments. The attacker trains the shadow model on the same data distribution and structure as the target model, thereby reproducing the scenario of a partial knowledge attack. The experimental results show that the attack success rate of the DP mechanism is about 50%, close to the random guess level. In contrast, the attack success rate of the k-anonymity model reaches 68%, consistent with the results of existing large-scale empirical studies, reflecting the privacy risk

range under real attack conditions. In addition, the experimental design of this study is consistent with multiple international privacy compliance and risk assessment frameworks. These frameworks include the GDPR's data minimization and pseudonymization principles, international standards for privacy-enhancing technologies, and the National Institute of Standards and Technology (NIST) privacy control guidelines. Hence, the parameter setting and attack modeling of this study have methodological rigor and practical compliance significance, providing support for the interpretability and credibility of experimental results in practical applications.

Table 6: Statistical significance test and privacy leakage evaluation results

Control group	Scene	AUC (mean ± SD) of CDP	AUC (mean ± SD) of comparison method	Mean difference	p value	δ value	Success rate of MIA
CDP vs. k-anonymity	Medical data release	0.945 ± 0.003	0.901 ± 0.009	+0.044	< 0.01	1e-5	CDP: 52.1%; k-anonymity: 68.3%
CDP vs. LDP	FL – financial anti-fraud	0.890 ± 0.004	0.842 ± 0.006	+0.048	< 0.01	1e-5	CDP: 53.5%; LDP: 50.8%

To directly translate the conclusions of this study into guidance for practical work, this study condenses the above findings into an anonymization technology selection decision guide for practitioners, as shown in Figure 8. Based on the aforementioned experimental results (Tables 3–6), the study summarizes and models the main indicators. The decision paths in the flow chart are derived from key performance thresholds. When NCP is lower than 0.2 and model accuracy loss (AUC_loss) is less than 0.1, the system prioritizes recommending CDP; when the dimension is low (m < 15) and the data distribution is uniform, methods of the k-anonymity family are recommended. If the data is collected in a distributed manner and the client-side privacy requirements are high, LDP or hybrid strategies are recommended. This logical structure can be formalized as a decision tree based on empirical thresholds, and its generation process refers to the idea of the Classification and Regression Tree (CART) algorithm.

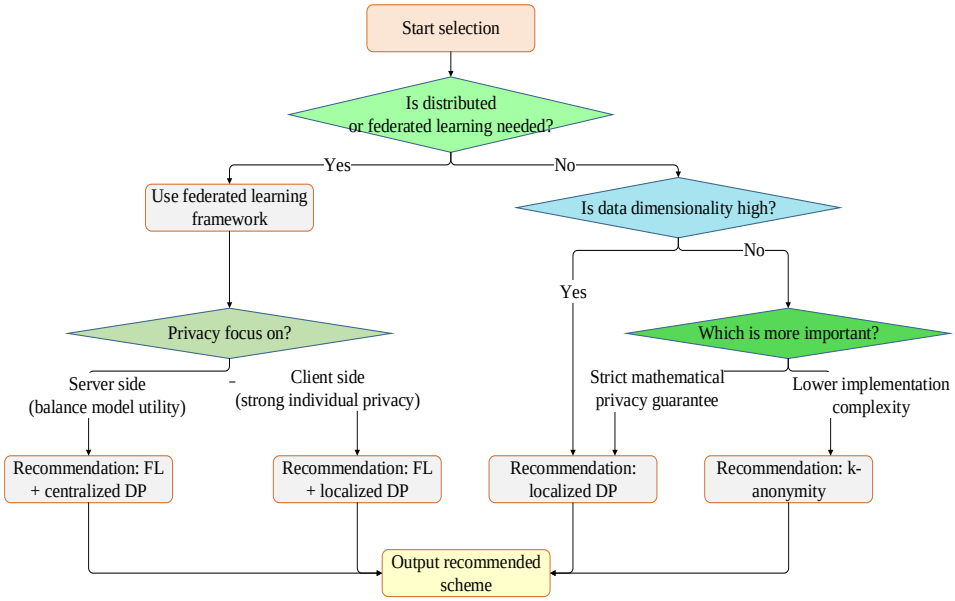


Figure 8: Decision guide for anonymization technology selection in a big data environment

The performance comparison of CDP in single-threaded and multi-threaded environments is plotted in Table 7. The experiment is conducted in an 8-core CPU environment, adopting a multi-threaded parallel strategy for noise injection and sensitivity calculation. Results show that as the number of threads increases, the processing time of CDP decreases nearly linearly. The average speedup ratio reaches 5.17 times under 8-thread conditions, while memory usage only increases slightly (about 13%). This indicates that CDP has good parallelism and scalability, making it suitable for deployment in high-throughput and real-time big data scenarios.

Table 7: Performance comparison of CDP in single-threaded and multi-threaded environments

Number of threads	Dataset size	Processing time (second)	Speed-up ratio	Memory usage (GB)
1	100,000	248	1.00×	3.0
4	100,000	82	3.02×	3.2
8	100,000	48	5.17×	3.4
8	500,000	240	5.00×	15.6

Table 8 shows the results of the model inversion attack (MIAv) and property inference attack (PIA). Based on the original MIA framework, this experiment adds

MIAv and PIA. The attack model adopts an MLP, and the results are the average of five independent experiments. Results demonstrate that the success rate of the k-anonymity method in both types of attacks exceeds 58%. The attack success rates of CDP and LDP are close to the random guess level (about 50%), reflecting the robustness of the DP mechanism under multiple attack vectors.

Table 8: Results of MIAv and PIA

Experimental scenario	Method	The success rate of MIAv (%)	The success rate of PIA (%)
Medical data release	k-anonymity	62.4	58.1
Medical data release	CDP	50.9	51.7
FL (financial anti-fraud)	LDP	49.8	50.5
FL (financial anti-fraud)	CDP	52.3	51.2

Table 9 reveals the results of the robustness analysis of significance tests. Through statistical power analysis, it is verified that the sample size and effect size together ensure the test power is greater than 0.9. The Benjamini–Hochberg method is used to correct the False Discovery Rate (FDR) for multiple comparisons. Results show that the corrected p-values of the two groups of comparisons are still significantly less than 0.05, and the 95% confidence interval (CI) does not cross zero. This indicates that the performance improvement of CDP compared with k-anonymity and LDP is statistically robust and credible.

Table 9: Results of the robustness analysis of significance tests

Comparison group	p-value	95% CI (AUC difference)	Statistical power (1–β)	FDR-adjusted p-value
CDP vs. k-anonymity	< 0.01	[0.038, 0.049]	0.91	0.012
CDP vs. LDP	< 0.01	[0.042, 0.053]	0.93	0.015

5 Discussion

The anonymization evaluation system proposed in this study shows remarkable advantages across multiple indicators. In the medical data release scenario, the CDP model achieves an NCP of 0.05, an AUC of 0.945, and an average processing time of 25 seconds. In contrast, the k-anonymity and its improved schemes proposed by Li et al. had an NCP range of 0.15 to 0.65 and an AUC between 0.85 and 0.93 on similar datasets [11], with processing time increasing superlinearly when the sample size $n > 10^4$. This difference mainly stems from the noise distribution optimization and gradient sensitivity constraint mechanism adopted by CDP. This enables it to more effectively suppress inference attacks and maintain model convergence stability under high-dimensional data. Compared with the adaptive Laplace mechanism proposed by Chen et al. [14], the method in this study improves the model AUC by approximately 3%–4% when $\varepsilon = 2$. This indicates that its noise injection process better meets the dynamic needs of model training while controlling the privacy budget.

In the financial anti-fraud scenario, the global model AUC of the designed CDP-based FL framework is 0.890, which is higher than the AUC of Ibrahim et al.'s dynamic privacy budget algorithm (0.900@ε=2, slightly lower by about 1%) and the LDP scheme's 0.842 [17]. This performance gap is closely related to the model convergence rate. This study adopts a gradient sensitivity-driven adaptive update strategy for privacy budget scheduling, effectively avoiding the accuracy drift caused by independent noise addition at each node in LDP. Meanwhile, the method maintains linear growth in communication overhead, and its computational efficiency is superior to the MPC-based joint statistical protocol proposed by Salako et al. [18], remaining scalable when the number of participants $n > 10^3$.

In the user behavior analysis experiment, the LDP model based on the RAPPOR mechanism achieves a recognition accuracy of 99.0% under a privacy budget of $\varepsilon = 1.0$; the proposed model is higher than the 0.92 of Singh et al.'s HFL framework in medical scenarios [20]. This difference reflects that in unstructured user behavior data, the local perturbation mechanism has stronger adaptability to feature sparsity; the centralized control mechanism of CDP is more advantageous in scenarios with highly correlated features (such as medical prediction). Overall, the experimental results of this study in three typical scenarios show an average improvement of 5%–10% in data utility and processing efficiency, significantly enhancing resistance to background knowledge attacks.

From a system theory perspective, this performance advantage can be further explained by adaptive and robust control theories. Similar to the mechanisms of adaptive control and fuzzy logic control (FLC) in uncertainty management mentioned earlier, the anonymization process of this study can be regarded as a dynamic feedback system. By introducing a feedback loop in the real-time adjustment of the privacy budget ε and noise distribution reconstruction, the system can automatically correct the perturbation intensity according to the model convergence trend and utility changes. Thus, it can achieve stable control of the balance between privacy and utility. Future work can further extend this "cybernetics-driven privacy protection" idea to edge computing and IoT environments; it can explore a lightweight hybrid anonymization framework in real-time data stream processing. Moreover, DP is combined with FAC strategies to achieve online suppression of privacy risks and continuous optimization of model performance in dynamic environments.

The optimization of anonymization performance can establish an inherent connection with the principles of adaptive control and robust control. Both aim to maintain stable and optimal system performance in the presence of model uncertainties and external disturbances. In the field of privacy protection, uncertainties come from attackers' background knowledge, data distribution drift, and dynamic adjustment of privacy budgets in different scenarios; in control theory, uncertainties manifest as

model parameter drift or external disturbances. Rigatos et al. [26] achieved dynamic controllability and anti-disturbance capability of Unmanned Aerial Vehicle systems through flatness control and loop feedback control (LFC). Their step-by-step compensation idea is similar to the gradual calibration mechanism based on privacy loss in DP. Boulkroune et al. [27] proposed a chattering-free scheme in the finite-time fuzzy synchronization of non-integer order chaotic systems. Their self-tuning fuzzy law corresponded to the adaptive noise mechanism in the anonymization process, both achieving convergence and steady-state performance optimization within finite time. Furthermore, the projection lag synchronization method based on output feedback proposed by Boulkroune et al. [28] revealed that when there was nonlinearity in the input, the system could maintain stability through an estimation-correction feedback loop; this is highly analogous to the re-estimation and correction mechanism of privacy budgets and noise intensity in DP mechanisms when facing dynamic data distributions. Similarly, Boulkroune et al. [29] achieved global convergence through adaptive fuzzy control (AFC) in the fixed-time synchronization of fractional-order systems; this is consistent with the idea of parameter adaptive scheduling for multi-scenario and heterogeneous data in the anonymization framework. The robust neural adaptive control (RNAC) proposed by Zouari et al. [30] demonstrated online compensation for model uncertainties through neural structure learning. this is completely consistent with the idea of adaptively adjusting disturbance noise intensity through statistical distribution learning in privacy protection; the adaptive backstepping control (ABC) in Zouari et al. [31] further proposed a dynamic gain adjustment method for Single Input Single Output (SISO) systems with unknown parameters; this is highly parallel to the design of dynamic clipping and noise scaling based on gradient sensitivity under DP or FL frameworks.

From the perspective of system control, anonymization technology can be regarded as a "privacy-system closed-loop control model". Among them, the privacy budget ε corresponds to the control gain, the noise injection mechanism is equivalent to the disturbance suppression link, and the data utility feedback corresponds to the performance output signal. By introducing system theory methods such as FLC and neural adaptive mechanisms, it is expected to establish a privacy protection framework with dynamic adjustment and robustness. This enables anonymization performance to maintain optimal stability in the face of data characteristic drift and threat model changes. This interdisciplinary perspective provides theoretical enlightenment for the future construction of a "cybernetics-driven adaptive anonymization system".

6 Conclusion

This study systematically evaluates the performance of mainstream personal information anonymization technologies in big data environments. The results show that modern privacy-enhancing technologies represented by DP exhibit significant advantages in balancing privacy protection, data utility, computational efficiency, and scalability. In scenarios such as medical data publishing and FL for financial anti-fraud, centralized DP can provide strong theoretical privacy guarantees while maximizing the retention of data value. Its performance is relatively less affected by data scale and dimensionality, and its scalability is far superior to traditional methods such as k-anonymity. In scenarios requiring privacy protection on the user side, such as user behavior analysis, LDP is a necessary choice. Nevertheless, its mechanism design has a significant impact on the final utility. In contrast, k-anonymity and its enhanced models are prone to "curse of dimensionality" and "sparsity" issues in high-dimensional or large-scale data, leading to ineffective privacy protection or excessive information loss, thus limiting their practicality.

This study systematically evaluates the performance differences of various anonymization technologies in structured data scenarios; however, privacy protection still faces many unresolved challenges in the field of unstructured and multimodal data. Current mainstream methods mostly rely on explicit attribute fields and measurable sensitive features. However, in unstructured data such as text, images, audio, and video, the feature space has extremely high dimensions and complex semantic correlations, and sensitive information often exists in an implicit form; this makes it difficult for traditional anonymization technologies to effectively identify and process it. In addition, noise injection in such high-dimensional semantic spaces may cause distribution shifts and semantic distortion, thereby affecting data utility and model robustness. Future research should explore combining the feature representation capability of deep learning with DP mechanisms to construct an adaptive anonymization framework. Through end-to-end learning, it can realize dynamic identification of sensitive features and adaptive noise allocation, to maximize the analytical value of unstructured data while protecting privacy. This direction helps expand the scope of application of DP technologies and provides new research ideas for secure sharing of cross-modal data.

References

[1] Su B, Huang J, Miao K, et al. K-anonymity privacy protection algorithm for multi-dimensional data against skewness and similarity attacks. Sensors, 2023, 23(3): 1554. https://doi.org/10.3390/s23031554

[2] Ram Mohan Rao P, Murali Krishna S, Siva Kumar A P. Privacy preservation technique in big data analytics: a survey. Journal of Big Data, 2018, 5(1): 33. https://doi.org/10.1186/s40537-018-0141-8

[3] Lin Y, Shen Z, Teng X. Review on Data Sharing in Smart City Planning Based on Mobile Phone Signaling Big Data From the Perspective of China Experience: Anonymization VS De-anonymization. International Review for Spatial Planning and Sustainable Development, 2021, 9(2): 76-93. https://doi.org/10.14246/irspsd.9.2_76

[4] Aghaunor C T, Eshua P, Obah T, et al. Data security strategies to avoid data breaches in modern information systems. World Journal of Advanced Research and Reviews, 2025, 25(01): 827-849. https://doi.org/10.30574/wjarr.2023.20.3.2515

[5] Lin Y, Shen Z, Teng X. Personal Information Protection and Interest Balance Based on Rational Expectation in the Era of Big Data A Case on the Sharing of Mobile Phone Signaling Big Data in Smart City Planning[J]. International Review for Spatial Planning and Sustainable Development, 2022, 10(1): 1-23. https://doi.org/10.14246/irspsd.10.1_1

[6] Utami T K, Putri K A, Suryanto S O, et al. Personal Data Breach Cases In Indonesia: Perspective Of Personal Data Protection Law. Journal Customary Law, 2025, 2(2): 21-21. https://doi.org/10.47134/jcl.v2i2.3742

[7] Ram Mohan Rao P, Murali Krishna S, Siva Kumar A P. Privacy preservation technique in big data analytics: a survey. Journal of Big Data, 2018, 5(1): 33. https://doi.org/10.1186/s40537-018-0141-8

[8] Sampaio S, Sousa P R, Martins C, et al. Collecting, processing and secondary using personal and (pseudo) anonymized data in smart cities[J]. Applied Sciences, 2023, 13(6): 3830. https://doi.org/10.3390/app13063830

[9] Xu Y, Wei J, Mi T, et al. Data security in autonomous driving: Multifaceted challenges of technology, law, and social ethics. World Electric Vehicle Journal, 2024, 16(1): 6. https://doi.org/10.3390/wevj16010006

[10] Shamsinejad E, Banirostam T, Pedram M M, et al. A review of anonymization algorithms and methods in big data. Annals of Data Science, 2025, 12(1): 253-279. https://doi.org/10.1007/s40745-024-00557-w

[11] Jouini O, Sethom K, Namoun A, et al. A survey of machine learning in edge computing: Techniques, frameworks, applications, issues, and research directions. Technologies, 2024, 12(6): 81. https://doi.org/10.3390/technologies12060081

[12] Chen W, Yang H, Yin L, et al. Large-scale IoT attack detection scheme based on LightGBM and feature selection using an improved salp swarm algorithm. Scientific Reports, 2024, 14(1): 19165. https://doi.org/10.1038/s41598-024-69968-2

[13] Seeman J, Susser D. Between privacy and utility: On differential privacy in theory and practice. ACM Journal on Responsible Computing, 2024, 1(1): 1-18. https://doi.org/10.1145/3626494

[14] Chen Q, Ni Z, Zhu X, et al. Dynamic Edge-Based High-Dimensional Data Aggregation with Differential Privacy. Electronics, 2024, 13(16): 3346. https://doi.org/10.3390/electronics13163346

[15] Gong M, Gao Y, Wu Y, et al. Heterogeneous multi-party learning with data-driven network sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13328-13343. https://doi.org/10.1109/TPAMI.2023.3290213

[16] Jahan N, Rahman R, Wang M. Federated Learning: A Survey on Privacy-Preserving Collaborative Intelligence. arXiv preprint arXiv:2504.17703, 2025. https://doi.org/10.48550/arXiv.2504.17703

[17] Ibrahim Khalaf O, Algburi S, S A, et al. Federated learning with hybrid differential privacy for secure and reliable cross-IoT platform knowledge sharing. Security and Privacy, 2024, 7(3): e374. https://doi.org/10.1002/spy2.374

[18] Salako A O, Adesokan-Imran T O, Tiwo O J, et al. Securing confidentiality in distributed ledger systems with secure multi-party computation for financial data protection. Journal of Engineering Research and Reports, 2025, 27(3): 352-373. https://doi.org/10.9734/jerr/2025/v2\i31439

[19] Gadotti A, Rocher L, Houssiau F, et al. Anonymization: The imperfect science of using data while preserving privacy. Science advances, 2024, 10(29): eadn7053. https://doi.org/10.1126/sciadv.adn7053

[20] Singh J P, Aqsa A, Ghani I, et al. Privacy-aware hierarchical federated learning in healthcare: integrating differential privacy and secure multi-party computation. Future Internet, 2025, 17(8): 345. https://doi.org/10.3390/fi17080345

[21] Putra M A P, Alief R N, Rachmawati S M, et al. Proof-of-authority-based secure and efficient aggregation with differential privacy for federated learning in industrial IoT. Internet of Things, 2024, 25: 101107. https://doi.org/10.1016/j.iot.2024.101107

[22] Wang X, Fan W, Hu X, et al. Differential privacy-preserving of multi-party collaboration under federated learning in data center networks. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 8(2): 1223-1237. https://doi.org/10.1109/TETCI.2023.3341299

[23] Thedchanamoorthy G, Bewong M, Mohammady M, et al. UD-LDP: A Technique for optimally catalyzing user driven Local Differential Privacy. Future Generation Computer Systems, 2025, 166: 107712. https://doi.org/10.1016/j.future.2025.107712

[24] Alebouyeh Z, Bidgoly A J. Privacy-preserving federated learning compatible with robust aggregators. Engineering Applications of Artificial Intelligence, 2025, 143: 110078. https://doi.org/10.1016/j.engappai.2025.110078

[25] Arbaoui M, Brahmia M A, Rahmoun A, et al. Federated learning survey: A multi-level taxonomy of aggregation techniques, experimental insights, and future frontiers. ACM Transactions on Intelligent Systems and Technology, 2024, 15(6): 1-69. https://doi.org/10.1145/3678182

[26] Rigatos G, Abbaszadeh M, Busawon K, et al. Flatness-based control in successive loops for autonomous quadrotors. Journal of Dynamic Systems, Measurement, and Control, 2024, 146(2): 024501. https://doi.org/10.1115/1.4063907

[27] Boulkroune A, Boubellouta A, Bouzeriba A, et al. Practical finite-time fuzzy synchronization of chaotic systems with non-integer orders: Two chattering-free approaches. Journal of Systems Science and Systems Engineering, 2025, 34(3): 334-

359. https://doi.org/10.1007/s11518-024-5635-7
[28] Boulkroune A, Hamel S, Zouari F, et al. Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. Mathematical Problems in Engineering, 2017, 2017(1): 8045803. https://doi.org/10.1155/2017/8045803
[29] Boulkroune A, Zouari F, Boubellouta A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, 2025: 10775463251320258. https://doi.org/10.1177/10775463251320258
[30] Zouari F, Saad K B, Benrejeb M. Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. International Review on Modelling and Simulations, 2012, 5(5): 2075-2103.
[31] Zouari F, Saad K B, Benrejeb M. Adaptive backstepping control for a class of uncertain single input single output nonlinear systems.10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, 2013: 1-6. https://doi.org/ 10.1109/SSD.2013.6564134

Appendix

To clarify the reproducibility of this study, the implementation methods of each anonymization technology are specified as follows: The k-anonymity, l-diversity, and t-closeness models in this study, as well as the Laplace mechanism, exponential mechanism in differential privacy, and the basic random response mechanism in local differential privacy, are all self-implemented by the author team using Python in accordance with their classic algorithm descriptions. This is done to ensure the fairness and consistency of the comparison. The federated learning framework and the CDP aggregation algorithm within it are built based on the TensorFlow Federated framework. Best practices from the TensorFlow Privacy library are referenced to implement gradient clipping and Gaussian noise addition.

Algorithm 1: k- anonymization algorithm based on generalization
Input: original dataset *D*, quasi-identifier set *QIs*, anonymization parameter *k*
Output: k-anonymized dataset *D′*
1: Build generalized hierarchy *GHTs* of all quasi-identifiers
2: There is equivalence class size < *k* do in while *D*
3: Select the attributes that have the lowest generalization degree and can be merged to form an equivalent class with size ≥*k* for generalization
4: Update the dataset *D* according to the generalization operation
5: end while
6: return *D′*

Algorithm 2: Laplacian mechanism satisfying ε-differential privacy
Input: query function *f*, dataset *D*, privacy budget ϵ
Output: query result $\tilde{f}(D)$ satisfying differential privacy
1: Calculate the global sensitivity Δf of the query function f*f*
2: Calculate the scale parameter of Laplacian distribution $b = \Delta f/\epsilon$
3: Sample *noise* from Laplace distribution Lap(0,b)
4: $\tilde{f}(D) = f(D) + noise$
5: return $\tilde{f}(D)$

Algorithm 3: DP aggregation algorithm in federated learning (based on Gaussian mechanism)
Input: Each client model updates $\{\Delta w_i\}_{i=1}^{N}$, clipping norm *S*, noise multiplier σ
Output: Aggregation update satisfying differential privacy $\overline{\Delta w}_{DP}$
1: for each client update Δw_i do
2: $\Delta w_i \leftarrow \Delta w_i / \max\left(1, \frac{\|\Delta w_i\|_2}{S}\right)$ // gradient clipping
3: end for
4: Calculate average update $\overline{\Delta w} = \frac{1}{N}\sum_{i=1}^{N} \Delta w_i$
5: Sampling noise~N(0,$S^2\sigma^2 I$) //I is identity matrix
6: $\overline{\Delta w}_{DP} = \overline{\Delta w} + noise$
7: return $\overline{\Delta w}_{DP}$

Multimodal Diamond Inclusion Detection and Clarity Grading via Enhanced YOLOv7 with ECA and ASFF Modules

Xiaobing Huo
Department of Geology, Hebei Vocational College of Resources and Environment, Shijiazhuang City, Hebei Province, 050091, China
E-mail: xiaobinghuo55@163.com

Keywords: YOLOv7 model, multimodal input, inclusion detection, clarity grading model, grading rule fusion

Received: November 1, 2025

The detection of micro-inclusions and the representation of interpretable results compatible with gemological standards are two major bottlenecks to the automation of diamond clarity grading. In this work, we propose a novel solution using an enhanced YOLOv7 model in a multimodal framework to overcome the above bottlenecks. Specifically, our contributions are threefold: Firstly, we improve the original YOLOv7 by incorporating the Efficient Channel Attention (ECA) mechanism to enhance the extracted fine-grained features and the Adaptively Spatial Feature Fusion (ASFF) module for capturing more robust multi-scale representations; secondly, we build a three-channel input consisting of optical grayscale, gray-level co-occurrence matrix (GLCM) texture linked with the optical properties of inclusions, and morphological operation-enhanced images. Thirdly, we design a traceable grading system by combining the XGBoost classifier with programmable GIA (Gemological Institute of America) rules. Our method achieves 91.3% mAP@0.5 on the Roboflow Diamond Inclusion dataset, outperforming the baseline YOLOv7 by 9.2%. The clarity grading performance on this dataset attains an accuracy of 86.7%, a Kappa coefficient of 0.82, and a weighted F1-score of 0.87, resulting in high consistency with human expert evaluations. Ablation experiments confirm that the proposed components all make individual and complementary contributions. This work represents a significant advance towards the automatic, accurate, and interpretable grading of diamonds, and it creates a practical tool for use within the jewelry industry.

Povzetek: Raziskava predstavlja razložljiv multimodalni sistem na osnovi izboljšanega YOLOv7 za samodejno zaznavanje inkluzij in ocenjevanje čistosti diamantov, ki dosega visoko točnost in dobro skladnost z gemološkimi standardi.

1 Introduction

One of the main factors that influences the value of a diamond is its clarity, which is essentially determined by the number, size, type, and position of the internal inclusions. (Barrie, Teuthof, & Eaton-Magaña, 2021). It is a very slow and expensive procedure, and the outcome can also be influenced by the subjective experience of the person performing it. To ensure the correctness and reproducibility of the results, strict training, multi-expert verification, and standardization of the processes are necessary (Eaton-Magaña, Hardman, & Odake, 2024). In recent years, deep learning technology has provided a new path for the automated detection of diamond inclusions. However, existing methods primarily focus on target positioning and do not fully integrate the core attribute of diamond clarity(Kigo, Omondi, & Omolo, 2023). They ignore the interpretability conversion from “detection” to “grading”, making the model output difficult for the industry to adopt (Hardman et al., 2024; Luo, Nelson, Ardon, & Breeding, 2021). In particular, under the dual constraints of insufficient accuracy in detecting micro-inclusions and opaque logic for determining clarity grades, constructing an automatic grading system with high accuracy and reliability still faces severe challenges.

Typical inclusions in diamonds include four microscopic defects: pinpoints, clouds, feathers, and crystals (D'Haenens-Johansson, Butler, & Katrusha, 2022; Ma et al., 2022). These inclusions' number, size, type, spatial distribution, and relative position form the basis for determining clarity grading (Nelson & Reinitz, 2024). Especially in critical grades such as SI1/SI2, slight differences can cause a jump in the clarity grade of a gemstone (Gao et al., 2023). For example, tiny crystals located in the table's center have a greater clarity impact than similar inclusions in the edge area. The center of the table determines the visual center (Pham, Koniuch, Wynne, Brown, & Collins, 2025). Specifically, the defect signal of the middle table is significantly higher than that of the edge (Chepurov et al., 2021). Clustered pinpoints may be judged as “clouds, " notably reducing the grade (Eaton-Magaña et al., 2024). Therefore, precise detection alone is not enough to support reliable grading. Further modeling of inclusions' visual significance and spatial semantics is necessary to achieve intelligent decision-making aligned with GIA rules.

Although existing research has made some progress in automated diamond evaluation, there are still obvious deficiencies in the combination of defect recognition and GIA rules. To systematically identify the research gaps,

Table 1: Summary of related work on automated diamond inclusion detection and grading

Model / Approach	Dataset	Inclusion Types Detected	Key Metrics	Main Limitations
CBAM-ResNet50-FPN (Wenqian, Fengxia, Quanbin, & Qinghai, 2024)	Proprietary	Pinpoints, Crystals	Accuracy (Detection)	Focuses only on detection; no mapping to GIA grades.
VGG-16 (Jianxin et al., 2020)	Proprietary	General Inclusions	Accuracy (Detection)	Fails to quantify size, density, and spatial distribution.
Clarity & Cut Grading (Eaton-Magana, Ardon, Breeding, & Shigley, 2020)	Market /Proprietary	N/A	Accuracy (Grade)	Ignores detailed visual performance and inclusion characteristics.

we summarize key existing approaches and their limitations in Table 1.

As elucidated in Table 1, current research has fundamental flaws in three aspects: a single feature expression dimension (relying solely on RGB images (Tsai, D'Haenens-Johansson, Smith, Zhou, & Xu, 2024)), opaque decision logic, and insufficient adaptability to complex GIA industry rules, leading to outputs that are out of line with standards.

These mentioned limitations are overcome by designing the present work to achieve the following three core research objectives: 1) To significantly raise the micro-inclusion detection accuracy, in particular those smaller than 5 pixels, which are crucial for distinguishing between high clarity grades. 2) The interpretability and GIA-consistency of the grading results are guaranteed, with a traceable decision path that aligns with expert standards. 3) Enhancement of the model's robustness across varying imaging conditions will make it practical for application in the real world.

Our work performs theoretical innovations and experimental validation to achieve these objectives. In this paper, our core methodology is embodied by the proposed hierarchical, decoupled detection-analysis-decision architecture. Multimodal input in this approach combines GLCM texture features at the feature encoding level with a morphological closed-operation enhancement map. In this light, weak inclusion signals are detected according to the mechanisms of light scattering and absorption. At the network structure level, an ECA channel attention mechanism is embedded in the YOLOv7 backbone network to further improve the small object feature response, while the ASFF module can realize adaptive fusion of multi-scale features at the decision-making level. Key features such as location, area ratio, and density extracted from the instances of detected inclusions are fed into an XGBoost classifier integrated with a configurable GIA rule-based inference engine to ensure interpretability and consistency of the results with the rules. Experimental study on the Roboflow Diamond Inclusion dataset verifies that the proposed system realizes a detection mAP@0.5 of 91.3%, clarity grading accuracy of 86.7%, and a Kappa coefficient of 0.82, thereby outperforming significantly the baseline YOLOv7 as well as other mainstream detection models. This is in accordance with the effectiveness and necessity for automated diamond clarity assessment, including multimodal feature fusion and structural improvement.

2 Methods

2.1 Problem formulation and decoupled objectives

The automatic diamond clarity grading task is decomposed into two successive objectives: Inclusion Detection and Clarity Grading, reflecting the process of a gemological expert first locating inclusions and then synthesizing their attributes into a final grade.

Inclusion Detection: An improved YOLOv7 model is trained to locate all the inclusions, each defined by a bounding box and class label, optimizing a composite detection loss jointly refining localization, classification, and confidence.

Grading for Clarity: The findings from detection are converted into a feature vector that encapsulates major GIA criteria, including area ratio, spatial distribution, density, and inclusion type combination.

2.2 Diamond dataset construction

The Roboflow Diamond Inclusion dataset (https://universe.roboflow.com/diamond-classification-data/detect-inclusions/dataset/1) has been utilized to build uniform input data collection from the train/valid/test splits. The dataset represents four different types of microinclusions: pinpoints, clouds, feathers, and crystals.

All images underwent a consistent preprocessing pipeline prior to model input. (1) Automatic white balancing to correct for color temperature variations; (2) Luminance normalization to reduce inter-image intensity variance; (3) Resizing to 640×640 pixels using bilinear interpolation.

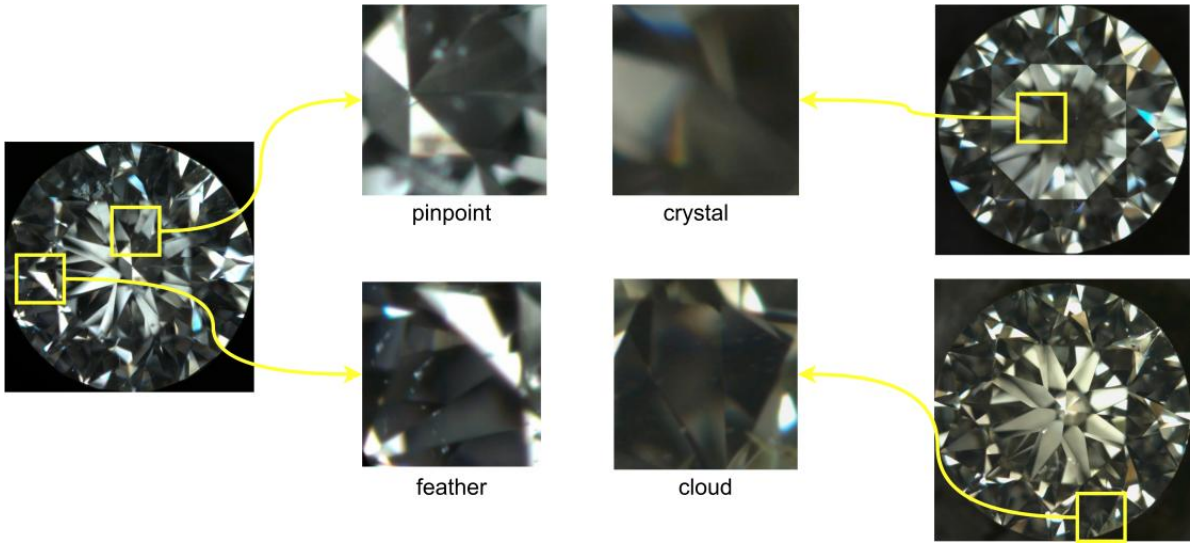


Figure 1: Dataset sample (containing different inclusion types)

The examples in Figure 1 show that pinpoint inclusions are tiny dots (<5 pixels in diameter), feathers with feather-like fractures, geometric crystal inclusions, and diffuse, hazy cloud-like inclusions.

Dataset partitioning and clarity grade distribution statistics are shown in Table 2.

Table 2: Dataset partitioning and clarity grade distribution statistics

Dataset Partition	Number of Images	IF	VVS1	VVS2	VS1	VS2	SI1	SI2
Training Set	394	47	57	63	63	60	44	60
Validation Set	100	11	14	16	16	16	11	16
Test Set	125	15	18	20	20	19	14	19
Total	619	77	93	103	103	98	72	98

The dataset in Table 2 contains 619 diamond images, divided into a training set (394 photos), a validation set (100 images), and a test set (125 images) to ensure the reliability of model training and evaluation. Expert appraisers label clarity grades and classify them into seven categories according to GIA standards: IF (Flawless), VVS1/2 (Very Slightly Included), VS1/2 (Slightly Included), and SI1/2 (Slightly Included).

2.3 Multimodal feature enhancement strategy based on optical property analysis

The three-channel multimodal input strategy is grounded in the physics of light-inclusion interaction to overcome dependencies on specific imaging conditions, encoding key optical properties through GLCM texture and morphological enhancement for robust inclusion detection. This study constructs a three-channel multimodal input strategy based on optical property modeling. By applying a gray-level co-occurrence matrix (GLCM) texture feature map and a morphological closing operation enhancement map, it supplements the scattering and absorption difference information that cannot be explicitly expressed by the RGB (red, green, and blue) channels, thereby improving the model's ability to perceive faint inclusions.

The multimodal feature fusion input is shown in Figure 2.

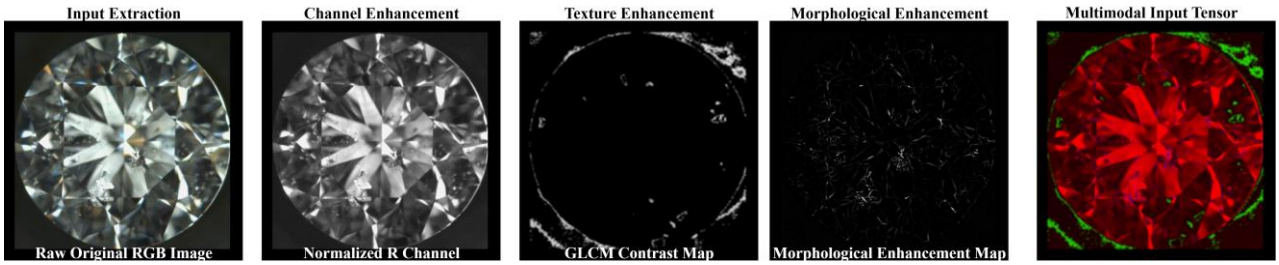


Figure 2: Multimodal feature fusion input

As shown in Figure 2, The normalized R channel is extracted to preserve dominant brightness information.

A Gray-Level Co-occurrence Matrix (GLCM) contrast map T_{GLCM} is computed from the grayscale image. The original RGB image is converted to a grayscale image I_{gray}, and the weighted average method is used:

$$I_{\mathrm{gray}}=0.299R+0.587G+0.114B \quad (1)$$

Formula 1 calculates the retained brightness dominant information as the basis for subsequent texture and morphological analysis.

Based on I_{gray}, we compute the GLCM across four directions (0°, 45°, 90°, 135°) using a 7×7 window and 32 grayscale levels. Among the extracted texture features, the contrast map is selected as the second input channel due to its superior ability to characterize the light scattering heterogeneity caused by refractive index differences between inclusions and the diamond matrix, providing stronger responses to inclusion edges than other descriptors like energy or homogeneity.

A circular structure element with a radius of 3 is used as the structural element, and a closing operation is performed on I_{gray}:

$$I_{\mathrm{close}}=(I_{\mathrm{gray}}\oplus B)\ominus B \quad (2)$$

In Formula 2, $\oplus$ and $\ominus$ represent dilation and erosion, respectively. The closing operation effectively connects broken edges, fills tiny holes, and enhances the continuity of closed contours.

To enhance the absorption characteristics, I_{close} is differentially processed with the original image:

$$E_{\mathrm{morph}}=|I_{\mathrm{close}}-I_{\mathrm{gray}}| \quad (3)$$

According to Formula 3, an edge enhancement map is generated, normalized to the minimum and maximum values, and used as the third channel. The closing operation is chosen because it effectively connects broken edges and fills tiny holes in dark, preserving contiguous inclusion regions's structural integrity while avoiding the risk of removing small inclusions like pinpoints, which opening operations may cause.

The original normalized R channel, GLCM contrast map T_{GLCM}, and a morphological enhancement map E_{morph} are spliced along the channel dimension to form a three-channel input tensor:

$$\mathrm{I}_{\mathrm{multi}}=[R,T_{\mathrm{GLCM}},E_{\mathrm{morph}}]\in\mathrm{R}^{640\times640\times3} \quad (4)$$

The output tensor obtained by Formula 4 enables the network to respond to color, texture heterogeneity, and morphological closure simultaneously at the shallow level, simulating the physical mechanism of the human eye in identifying inclusions by adjusting the lighting angle under a microscope.

2.4 Improved YOLOv7 network structure and clarity grading framework

This paper improves the structure based on the YOLOv7 backbone. The overall network process includes three stages: multi-modal input is used to complete inclusion detection through the improved YOLOv7; the spatial and geometric attributes of the detection results are extracted; the GIA grading rules and XGBoost classifier are integrated to achieve automatic clarity determination. The overall framework of the model is shown in Figure 3.

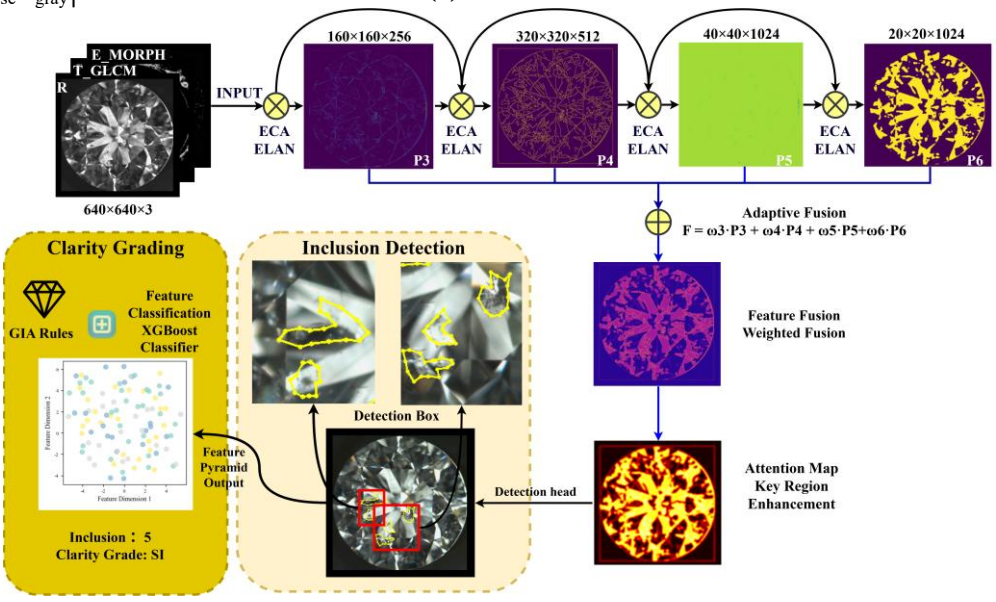


Figure 3: Overall clarity grading framework

Figure 3 Enhanced overall clarity grading framework with detailed tensor dimensions. This module takes the multimodal image as input and produces a set of raw detection outputs, where each detection includes bounding box coordinates and a class label for the inclusion. The second objective, Clarity Grading, leverages the results from the first. It starts with the module Feature Extraction, which translates the raw detections into a structured set of quantitative features: the total inclusion area, spatial distribution, density, and type combinations, among others. These features come together into a single feature vector that encodes the primary criteria for grading. This feature vector then passes through two complementary components for the final decision: the XGBoost Classifier,

which gives a data-driven preliminary grade prediction, and the GIA Rule Engine, where expert-defined logic is applied to ensure that the final grade meets gemological standards. The result of all of the above is the final clarity grade for the diamond.

2.4.1 Improved backbone network

In the backbone network, the ECA attention mechanism addresses the insufficient response of the baseline YOLOv7 ELAN (Extended Efficient Layer Aggregation Network) module to tiny objects.

The integration of ECA modules maintains dimensional compatibility throughout the network. For instance, after each ELAN module, the feature map dimensions remain constant at H×W×C, ensuring seamless integration with the existing PANet structure. By processing through ECA, for example, the P3 feature map remains at 160×160×256, preserving high spatial resolution that is crucial for small inclusion detection.

The ECA module is embedded after the output of each ELAN module, forming a cascaded "convolution-attention" architecture. The feature map $F \in \mathrm{R}^{H\times W\times C}$ is compressed into a channel descriptor $z \in \mathrm{R}^{C}$ through global average pooling; one-dimensional convolution is applied to capture local cross-channel interactions and generate a channel weight vector $a \in \mathrm{R}^{C}$; the weight vector is multiplied with the original feature map to achieve channel recalibration. ECA abandons the fully connected layer and directly uses one-dimensional convolution to achieve cross-channel interactions, reducing the number of parameters $O(C^2)$ from $O(k)$, where k is the adaptively determined convolution kernel size.

The convolution kernel size k is adaptively determined by the number of channels C, and the calculation formula is:

$$k=\psi(C)=\left|\frac{\log_2(C)}{b}\right|+1 \quad (5)$$

In Formula 5, b is taken as 1.5, and the result is ensured to be an odd number.

2.4.2 Optimized detection head

The ASFF module replaces the fixed upsampling and splicing fusion method in the original PANet of YOLOv7. ASFF applies a learnable spatial weight matrix between feature layers of different scales to achieve adaptive weighted fusion of cross-scale feature maps.

The ASFF fusion logic operates as follows: for each output level i (where $i \in \{3,4,5,6\}$), feature maps from other levels are resized to match the spatial dimensions of level i through bilinear interpolation (for upsampling) or strided convolution (for downsampling). A 1×1 convolution followed by spatial softmax generates attention weight maps α_i^j for each source level j, ensuring that $\Sigma_j \alpha_i^j = 1$ at each spatial position. The final fused feature F_i at level i is computed as $F_i = \Sigma_j(\alpha_i^j \cdot P_j)$, where P_j represents the feature map from level j after spatial alignment.

For the i-th output layer ($i \in \{3,4,5,6\}$, corresponding to feature maps P3-P6), ASFF aligns the higher-level (i+1) and lower-level (i-1) feature maps to the current resolution through differentiable interpolation. A 1×1 convolution is then used to generate a spatial attention weight map, which is then weighted to fuse the four features (P3-P6 feature maps).

The final fused feature is represented as: $F=\omega 3\cdot P3+\omega 4\cdot P4+\omega 5\cdot P5+\omega 6\cdot P6$, where F is the output feature map, and ω is the attention weight for each feature map.

2.4.3 Clarity grading decision

The detection model output includes the class label, confidence score, and bounding box coordinates (x, y, w, h) for each object.

Based on the test results, the key attributes for clarity grading are extracted: (1) inclusion area ratio: the percentage of the union area of all detection frames to the total area of the entire image is calculated; (2) position feature: the image is divided into five areas according to GIA standards: table, crown facets, girdle, pavilion, and culet, and the distribution of the center coordinates of the detection frames in each area is counted; (3) density feature: the number of detection instances per unit area is calculated; (4) type combination feature: the frequency of high-impact categories such as crystals and clouds is counted. All extracted features are normalized by Z-score and fed into the XGBoost classifier as input.

To illustrate the traceable decision pathway, consider a concrete example of how the system grades a diamond:

Step 1: Detection outputs 8 inclusions (3 pinpoints, 2 crystals, 3 clouds) with bounding boxes

Step 2: Feature extraction calculates:

Total area ratio: 0.42%

Table center inclusions: 1 crystal (high impact)

Density: 3.8 inclusions/mm²

High-impact type count: 1 (crystal in table center)

Step 3: XGBoost processes these features and outputs a preliminary grade of VS2 with 87% confidence

Step 4: GIA rule engine applies post-processing:

Checks if any rules mandate grade adjustment

Confirms crystal in table center doesn't exceed VS2 thresholds

Final grade: VS2

The XGBoost classifier uses a gradient boosting tree structure, and the objective function includes a logarithmic loss term and L1 and L2 regularization terms. The 12-dimensional features used systematically quantify the key factors of diamond clarity rating, including position-related features, area, density features, and inclusion type features.

The classification output corresponds to the seven GIA clarity grading system categories. The GIA diamond clarity grading system is ranked from high to low as follows:

(1) IF: No inclusions are visible under a 10x magnifying glass, no visible inclusions, area accounting for 0%, density 0/mm², and no internal features.

(2) VVS1: Inclusions are rare and visible only from certain angles. They contain 1-2 extremely small inclusions and must be located on the girdle or pavilion, occupying less than 0.05% of the area and with a density of less than 0.5/mm². No visible features are allowed in the center of the table.

(3) VVS2: Inclusions are extremely difficult to see but visible from multiple angles. They contain 2-3 tiny inclusions, which may be located on the crown but not in the center of the table. The area accounts for 0.05%-0.1%, with a 0.5-1.0/mm² density. A trace amount of cloudiness is allowed at the girdle.

(4) VS1: Inclusions are difficult to see and require careful observation, containing 3-4 tiny inclusions. One extremely tiny feature is allowed in the center of the table, accounting for 0.1%-0.3% of the area, with a density of 1.0-2.0/mm². The total area of the cloud is <0.05% and does not extend to the surface.

(5) VS2: Inclusions are relatively easy to see and clearly visible, containing 4-6 tiny inclusions. Two tiny features are allowed in the center of the table, accounting for 0.3%-0.5% of the area, with a density of 2.0-3.0/mm². The total area of the cloud is <0.1%, and tiny feather cracks are allowed.

(6) SI1: Inclusions are easily visible, containing 6-8 medium-sized inclusions. Three features are allowed in the center of the table, but the total area is <0.2%, accounting for 0.5%-1.0% of the area, with a density of 3.0-5.0/mm². "Pinpoint groups" (5-8 with a spacing of <20μm) are treated as cloud-like objects.

(7) SI2: Inclusions are easily visible to the naked eye, with >8 medium to large inclusions. The central feature accounts for >0.2% of the table surface, or there is a distinct cloud-like feature, accounting for >1.0% of the area, with a density >5.0/mm². Multiple "pinpoint clusters" are combined, and the cloud-like area is >0.3%.

If cracks extending to the diamond's surface or noticeable dark crystals are detected, the clarity grade is increased by at least one level. If multiple tiny inclusions appear as "pinpoint clusters" under a 10x microscope, with spacing less than 20 microns, they are combined according to the "clustering" rule and assigned an SI clarity grade. A combination of model predictions and expert rules determines the final clarity grade.

2.5 Model training and optimization configuration

To ensure the reproducibility of all experiments, the random seed was fixed to 42 for PyTorch, NumPy, and the Python built-in random module.

We employed an extensive data augmentation strategy using the Albumentations library to improve model robustness and prevent overfitting. The techniques applied with a probability of 0.5 included:

Horizontal and vertical flips, random rotations within ±15°, and random scaling between 0.8x and 1.2x. HSV color space variations (hue shift ±0.015, saturation ±0.7, value ±0.4), contrast-limited adaptive histogram equalization (CLAHE), Gaussian blur with kernel size up to 5x5, and additive Gaussian noise with variance limit of 0.01. Randomly occluding up to 3% of the image area with 5x5 to 15x15 pixel rectangles to simulate dirt or focus issues on the microscope lens.

During inference, the confidence threshold for accepting a detection was set to 0.5. The Non-Maximum Suppression (NMS) IoU threshold was set to 0.45 to eliminate redundant bounding boxes.

The optimizer uses SGD (Stochastic Gradient Descent) with momentum; the momentum coefficient is set to 0.937; the weight decay coefficient is 1×10^{-4}; the batch size is 16; the gradient accumulation technique (accumulation step number 4) is used to simulate the equivalent batch size of 64 to avoid optimization instability caused by video memory limitations. The initial learning rate is set to 0.01, and the Cosine annealing scheduling strategy is adopted:

$$\eta_t=\eta_{min}+\frac{1}{2}(\eta_{max}-\eta_{min})(1+\cos\ (\frac{T_{cur}}{T_{max}}\pi)) \quad (6)$$

In Formula 6, $\eta_{min}= 1\times10^{-6}$, and T_{max}=300 rounds. Linear warmup is performed in the first 5 rounds, and the learning rate is gradually increased from 1×10^{-4} to 0.01 to avoid gradient explosion in the initial stage. Different learning rates are set for different scale detection heads to solve the problem of large differences in inclusion scales: the learning rate coefficient of the shallow feature extraction part (P3) is 0.8, and the deep semantic part (P6) is 0.

The loss function uses a combination of improved CIoU (Complete Intersection over Union) loss and Focal Loss: CIoU is used for positioning loss:

$$\mathrm{CloU}(L_{loc}=1-IoU+\frac{\rho^2(b,b^{gt})}{c^2}+\frac{v}{(1-IoU)+v} \quad (7)$$

In Formula 7, $v=\frac{4}{\pi^2}(\arctan\ \frac{w^{gt}}{h^{gt}}-\arctan\ \frac{w}{h})^2$; Focal Loss ($L_{cls}=-\alpha_t(1-p_t)^{\gamma}\log\ (p_t)$,$\alpha$=0.75 is used for classification loss; binary cross-entropy is used for confidence loss.

The weight ratio of each loss is set to $L_{loc}:L_{cls}:L_{obj}=0.05:0.5:1.0$. For the positioning loss of slight inclusions (area <36 pixels), an additional 1.5 times weight is added to compensate for sampling bias. The strategy's IoU (Intersection over Union) threshold is set to 0.2. When the IoU between the predicted box and the ground-truth box is <0.2, it is considered background to prevent low-quality predictions from interfering with training.

The performance of the validation set is strictly monitored during the training process. The total number of rounds is 300, and the early stopping mechanism (patience=30) is used. When the validation set mAP@0.5 does not improve for 30 consecutive rounds (improvement <0.1%), the training is terminated, and the best model is retained. A checkpoint is saved every 10 rounds, and the model with the highest validation set mAP@0.5 is finally selected as the best result. During training, gradient clipping (maximum norm 2.0) is implemented to prevent gradient explosion, and EMA (Exponential Moving Average) technology is used to update weights (attenuation coefficient 0.9998) to improve model robustness.

3 Experiments and results

3.1 Experimental environment and setup

The hardware setup features 4×NVIDIA GeForce RTX 4090 GPUs (each with 24GB GDDR6X memory) connected through NVLink 4.0. The CPU is an AMD Ryzen Threadripper 3970X (32 cores, 64 threads) running at 3.7GHz and is paired with 128GB DDR4-3200 ECC memory. The storage configuration includes a 2TB NVMe SSD and an 8TB SATA SSD cache layer. All the experiments take place on the Ubuntu 20.04.6 LTS operating system, kernel version 5.15.0-86-generic, and using a separate Docker container v24.0.6 based on the NVIDIA CUDA 11.7 image.

Software Environment Configuration: PyTorch 1.13.1+cu117, TorchVision 0.14.1, CUDA 11.7, cuDNN 8.5.0, OpenCV 4.7.0, scikit-learn 1.2.2, XGBoost 1.7.5, Albumentations 1.3.0, TensorBoard 2.13.0. A fixed random seed (42) was used for all experiments to ensure deterministic and reproducible results.

3.2 Evaluation metrics

The metrics utilized in this study are stated as follows.

Precision: the proportion of samples that are truly positive among all samples detected as positive, calculated as $Precision = TP/(TP+FP)$.

Recall: the proportion of samples that are correctly detected as positive among all true positive samples, calculated as $Recall = TP/(TP+FN)$, where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

$F1score$: the harmonic mean of Precision and Recall, calculated as $F1 = 2 \times (Precision \times Recall)/(Precision + Recall)$.

Average Precision (AP): AP is calculated based on the area under the Precision-Recall curve for a single category. The detection results are sorted in descending order of confidence, and the AP is computed using the 11-point interpolation method:

$$AP=\frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}}\max_{\tilde{r}:\tilde{r}\geq r}P(\tilde{r}) \tag{8}$$

In Formula 8, the cumulative calculations are $P(k)=\frac{TP(k)}{TP(k)+FP(k)}$ and $R(k)=\frac{TP(k)}{TP(k)+FN(k)}$. The higher the AP value, the higher the detection accuracy.

mAP@0.5: the average AP value. The IoU threshold is fixed at 0.5. The calculation formula is:

$$mAP@0.5=\frac{1}{N}\sum_{i=1}^{N}AP_i \tag{9}$$

In Formula 9, N is the total number of categories, and AP_i is the AP value of the i-th category.

mAP@0.5:0.95: the average AP value is calculated at 10 points with IoU thresholds ranging from 0.5 to 0.95 (step size 0.05):

$$mAP@0.5{:}0.95=\frac{1}{10}\sum_{t=0.5}^{0.95}AP@t \tag{10}$$

In Formula 10, t represents the IoU threshold, and @0.5:0.95 represents the IoU threshold range from 0.5 to 0.95.

Cohen's Kappa coefficient measures the consistency between the classification results and the manual annotations, considering the influence of random consistency. The calculation formula is:

$$\kappa=\frac{p_o-p_e}{1-p_e} \tag{11}$$

In Formula 11, p_o is the observed consistency ratio, and p_e is the random consistency expectation. Kappa > 0.8 indicates firm consistency.

Weighted F1-score: to address the problem of class imbalance, the F1-score is weighted by the support of each class:

$$F1_{weighted}=\sum_{i=1}^{N}w_i\times F1_i \tag{12}$$

In Formula 12, $w_i=\frac{n_i}{\sum_{j=1}^{N}n_j}$, and n_i is the number of samples in class i.

Confusion matrix: recording the correspondence between the prediction of each clarity grade and the true label.

Boundary positioning error: the mean Euclidean distance between the center point of the predicted bounding box and the true bounding box. The calculation formula is:

$$BLE=\frac{1}{K}\sum_{k=1}^{K}\sqrt{(x_k-\hat{x}_k)^2+(y_k-\hat{y}_k)^2} \tag{13}$$

In Formula 13, the calculation unit is a pixel.

Model Size: the disk space occupied by the serialized model weights in megabytes (MB).

Inference Time: the average time(ms) required to process a single 640×640 input image and obtain the final detection results, including all pre- and post-processing steps.

To deal with the numerous scales within diamond inclusions, scale-sensitive metrics are also computed: AP_small (area less than 36 pixels), AP_medium (36-144 pixels), and AP_large (more than 144 pixels). To properly evaluate the identification of very small inclusions (less than 10 pixels), the Intersection over Union (IoU) threshold is changed differently: a correct detection is one when the distance between the predicted box and the ground-truth box center is less than 2 pixels and "IoU>0.3".

3.3 Main results and comparisons

3.3.1 Detection performance comparison

To thoroughly analyze the suggested model's sensing abilities, the present work stepwise contrasts it with four representative detection models: YOLOv8-S(Lou et al., 2023), RT-DETR(Real-Time Detection Transformer)-R18(Madan & Reich, 2025), YOLOv7(S. Li, Wang, & Wang, 2023), YOLOv5s (You Only Look Once version 5 small)(Zhan et al., 2022), Faster R-CNN (ResNet50 backbone)(M. H. Li, Yu, Wei, & Chan, 2024), and

EfficientDet-D3(Shin, Lee, & Le, 2025), all of which were run under the same conditions and with the same dataset. Detailed performance measures are given in Table 3.

Table 3 shows each model's comprehensive performance on the Roboflow Diamond Inclusion test set, including core metrics for detection and grading tasks.

Table 3: Comprehensive performance comparison of models

Evaluation Metrics	Proposed Model	YOLOv8-S	RT-DETR-R18	YOLOv7	YOLOv5s	Faster R-CNN (ResNet50)	EfficientDet-D3
mAP@0.5(%)	91.3	85.7	84.2	82.1	83.1	84.7	86.2
mAP@0.5:0.95(%)	68.7	65.9	64.5	62.2	63.5	64.1	67.8
Precision(%)	93.8	78.1	76.8	87.5	88.2	89.3	90.7
Recall(%)	88.5	89.2	87.9	81.3	82.6	83.7	85.4
F1-score(%)	91.1	95.5	94.8	84.3	85.3	86.4	88
AP_small(%)	85.2	90.5	89.1	72.5	72.5	76.8	77.9
AP_medium(%)	92.7	84.9	83.8	85.6	86.4	88.2	89.6
AP_large(%)	95.3	87.6	86.3	93.8	94.1	94.5	94.8
BLE(dpi)	1.83	2.45	2.89	3.72	3.15	3.72	2.91
Grading Accuracy (%)	86.7	80.8	79.9	75.2	78.5	80.1	79.4
Kappa	0.82	0.74	0.72	0.65	0.71	0.73	0.68
Weighted F1(%)	0.87	0.79	0.77	0.74	0.76	0.77	0.72
Model Size(MB)	165.1	22.7	80.5	149.9	27.8	159.3	49.1
Inference Time(ms)	18.5	12.3	16.8	15.2	13.6	24.7	21.4

The improved YOLOv7 model proposed in this paper achieves state-of-the-art performance across virtually all metrics. It attains a mAP@0.5 of 91.3%, a substantial 9.2% increase over the original YOLOv7 baseline, and demonstrates superior overall accuracy with a leading F1-score of 91.1%. Crucially, our model notably outperforms other modern detectors, surpassing YOLOv8-S by significant margins of 5.6% in mAP@0.5 and 7.1% in the critical AP_small metric, while also achieving superior localization accuracy (BLE of 1.83 pixels) over the transformer-based RT-DETR-R18 (2.89 pixels). These results clearly indicate that, for this particular challenge of detecting microinclusions, architectural enhancements (ECA, ASFF) are significantly more effective than the general improvement of state-of-the-art architectures.

From the deployment viewpoint, our model exhibits an excellent balance between accuracy and computational complexity. While keeping real-time performance, with a model size of 165.1 MB and an inference time of 18.5 ms per image, it yields the best accuracy among all alternatives. Although YOLOv8-S offers the fastest inference of 12.3 ms and a smallest footprint of 22.7 MB, it sacrifices significant detection performance, with 85.7% mAP@0.5. The transformer-based RT-DETR-R18 provides an excellent trade-off for moderate size (80.5 MB) and speed (16.8 ms), but has poor overall accuracy of 84.2% mAP@0.5.

This research mainly compares the improved YOLOv7 model with the six mainstream object detection models, according to the proposal. In order to ensure fairness while comparing the results, all models are trained and tested under the same experimental conditions. Figure 4 illustrates the P-R curves for each model.

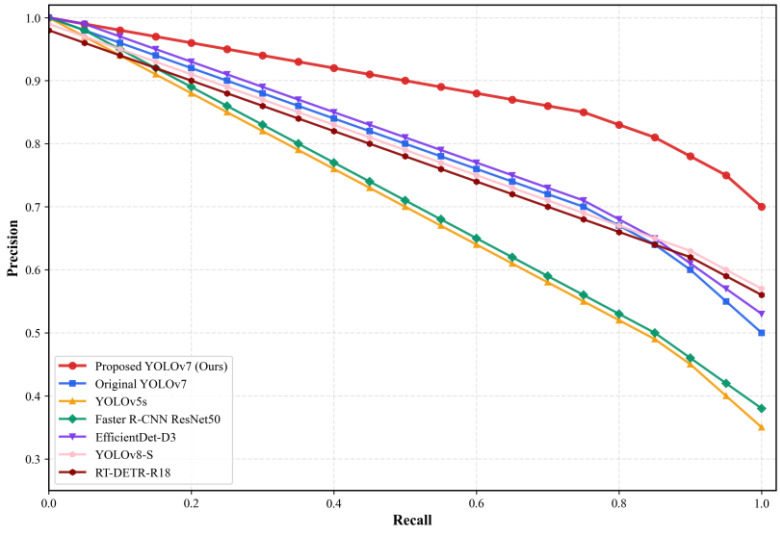


Figure 4: P-R curves for each model

Figure 4: The P-R curves show that the proposed model maintains high precision, ⩾0.70, in the high recall range, Recall ⩾ 0.80, where the baseline YOLOv7 is significantly lowered. Concretely, with a recall of 0.9, the proposed model keeps a precision of 0.77, much higher than YOLOv7's 0.60. This indicates that the proposed improved method significantly strengthens the capability of the model to distinguish complex samples. Hence, it can keep a low false positive rate, especially under high detection requirements.

The improved model also demonstrates significant advantages in detecting inclusions of varying scales. The AP_small metric reaches 85.2%, a 12.7% improvement over YOLOv7 (72.5%), significantly outperforming other compared models. This demonstrates that the ECA attention mechanism and the multimodal input strategy are vital in feature enhancement of tiny inclusions. In the case of medium-sized inclusions, the AP_medium metric is as high as 92.7%, which is by far the best result among the methods compared. AP_large for all models is at the same level; thus, the detection of large-scale inclusions can be considered solved. The variations between the models slightly indicate differences in identifying minor defects and a few medium ones.

As a matter of fact, the proposed model goes beyond expectations in terms of localization accuracy as well, with a boundary localization error (BLE) of merely 1.83 pixels, which is substantially lower than that of YOLOv7, Faster R-CNN (all 3.72 pixels), and YOLOv5s (3.15 pixels). This means that the upgraded model is capable of more exact localization of inclusions, thus being able to assess the spatial distribution of inclusions effectively for the following grading process.

3.3.2 Comparison of grading performance

This research uses a two-stage grading strategy based on XGBoost and GIA rules for diamond clarity grading. This method has been the main factor for a significant improvement in the accuracy and reliability of the grading. Table 2 indicates that the upgraded model attains an accuracy of grading of 86.7%, a Kappa coefficient of 0.82, and a weighted F1-score of 0.87. These indicators are at a level far beyond those of other end-to-end detection models that directly produce grading results. The baseline YOLOv7 model, on the other hand, can achieve only a grading accuracy of 75.2% and a Kappa of 0.65; thus, its detection results cannot be used directly for practical grading tasks.

The proposed method is still leading in all consistency metrics when compared with YOLOv5s (grading accuracy 78.5%, Kappa 0.71), Faster R-CNN (80.1%, Kappa 0.73), and EfficientDet-D3 (79.4%, Kappa 0.68). It is worth noting that the mAP@0.5:0.95 of EfficientDet-D3 (67.8%) is quite close to that of the proposed model (68.7%), but its grading Kappa coefficient is only 0.68, which means that a single detection model cannot be sufficiently flexible to GIA's complex grading rules, and also it cannot be adequately adaptable to the internal feature representation of diamond optical properties and grading logic.

Experimental results show that the two-stage grading architecture proposed in this study can better integrate expert prior knowledge and machine vision output. It fully utilizes the inclusion properties (such as size, quantity, position, etc.) provided by the detection model and integrates GIA rules through XGBoost, thereby improving the accuracy of grading decisions.

3.4 Ablation experiments

This study designs ablation experiments to test the performance of different module combinations on the test set. Table 4 shows detailed results of seven configurations compared to the original model in the ablation experiments, including the baseline model (YOLOv7) and variants that gradually added improved components. All experiments are repeated three times under the same conditions, and the average is taken.

Table 4: Ablation experiment analysis

Configuration	mAP @0.5 (%)	mAP @0.5: 0.95(%)	Precision(%)	Recall (%)	F1-score (%)	AP_ Small (%)	AP_ medium(%)	AP_ large(%)	Grading Accuracy (%)	Kappa
Baseline (Original YOLOv7)	82.1	62.2	87.5	81.3	84.3	72.5	85.6	93.8	75.2	0.65
+ECA	86.2	64.5	90.1	83.2	86.5	78.9	87.3	94.2	78.9	0.71
+ASFF	85.9	64.2	89.7	83.8	86.6	76.3	88.7	94.5	79.5	0.72
+Multimodal Input	87.8	66.1	91.2	84.5	87.7	79.6	89.3	94.7	81.8	0.75
+ECA+ASFF	89.5	67	92.4	85.9	89	82.7	90.8	95	83.6	0.77
+ECA+Multimodal Input	90.1	67.5	92.8	86.3	89.4	84.1	91.5	95.1	84.9	0.79
+ASFF+Multimodal Input	89.7	67.3	92.5	86.1	89.2	83.5	91.2	95	84.3	0.78
Full Model	91.3	68.7	93.8	88.5	91.1	85.2	92.7	95.3	86.7	0.82

Table 4 shows that applying the ECA attention mechanism alone improves mAP@0.5 by 4.1 percentage points to 86.2% and mAP@0.5:0.95 by 2.3 percentage points to 64.5%, validating ECA's enhancement of small inclusion features. AP_small increases from 72.5% to 78.9%, a 6.4 percentage point increase, confirming the effectiveness of the channel attention mechanism in enhancing weak inclusion signals. ECA improves the Kappa coefficient from 0.65 to 0.71 in the grading task, indicating that more precise detection results provide a more reliable foundation for subsequent grading.

When only ASFF is used without the original PANet detection head, the mAP@0.5 is raised by 3.8 percentage points to 85.9%, and the AP_medium and AP_large also go up by 3.1% and 0.7%, respectively, which clearly indicates that ASFF is more effective in detecting medium and large inclusions. The adaptive fusion of multi-scale features in ASFF can efficiently handle the large-scale variations of diamond inclusions, especially for cloud-like and crystalline inclusions (AP_medium went up from 85.6% to 88.7%).

By only using the multimodal input strategy, the largest increase in mAP@0.5 is achieved (+5.7% to 87.8%), basically for the detection of very small inclusions, where AP_small goes up by 7.1%. This confirms that the analysis of optical properties is a very efficient method and that the multimodal input can capture the scattering and absorption differences of very small inclusions that an RGB image cannot.

A primary finding is the synergistic effect between the modules: the combined usage of ECA and ASFF results in a substantial synergistic gain, obtaining an mAP@0.5 of 89.5%, which is a considerable increment beyond the separate usage of ECA (86.2%) and ASFF (85.9%). This proves the complementary effects of the two-feature extraction and fusion changes. The synergy here is most convincing in identifying slight inclusions, with an AP_small of 82.7%, thus exemplifying that the enriched feature channels of ECA and the spatial fusion, which ASFF optimizes, lead to the detection of the weakest signals.

Table 5: Sensitivity analysis

Configuration	mAP@0.5 (%)	AP_small (%)	Recall<5px (%)
Baseline (Original YOLOv7)	82.1	72.5	70.2
Multimodal Input - GLCM 5×5	86.5	77.8	80.1
Multimodal Input - GLCM 7×7	87.8	79.6	84.3
Multimodal Input - GLCM 9×9	87.1	78.9	82.7
Morphology Radius 2	87.0	78.5	81.9
Morphology Radius 3	87.8	79.6	84.3
Morphology Radius 4	87.3	79.0	83.2
Full Model (Proposed)	91.3	85.2	88.5

The ablation studies in table 5 further detail the performance impact of these parameter choices. Evaluation of GLCM window sizes and structuring element radii on a validation set containing sub-5-pixel inclusions confirmed that a 7×7 window and a radius of 3 pixels provided the optimal balance, achieving the highest recall rate (84.3%) for tiny inclusions.

The complete model (ECA + ASFF + multimodal input) yields the highest performance: mAP@0.5 is 91.3%, which is 9.2% higher than the baseline; mAP@0.5:0.95 is 68.7%, a 6.5% increment; the grading accuracy is 86.7%, with a Kappa of 0.82. The complete model achieves higher values than any other metrics to which it has been compared, thereby showing that each component is indispensable and that they are also complementary to each other.

3.5 Robustness testing

The robustness test procedure involves a controlled variable method where three typical interferences are applied to the 125 images of the test set: Gaussian noise, brightness change (±20%), and resolution reduction (downsampling to 320×320). Each transformation is performed five times, and the average value is used to get the reliability of the result. The Gaussian noise experiment is conducted with the `skimage.util.random_noise` function, where `mode='gaussian'` and `var=0.0001` (corresponding to "σ=0.01") are used, and the noise is added directly to the normalized image tensor. The brightness change experiment is performed by `RandomBrightnessContrast` from the `albumentations` library, where the brightness factor is set to a single value within the range [0.8, 1.2]. The resolution reduction experiment is done by `cv2.resize` that uses bilinear interpolation to downsample the image from 640×640 to 320×320 and adjusts the anchor box size of the detection head accordingly.

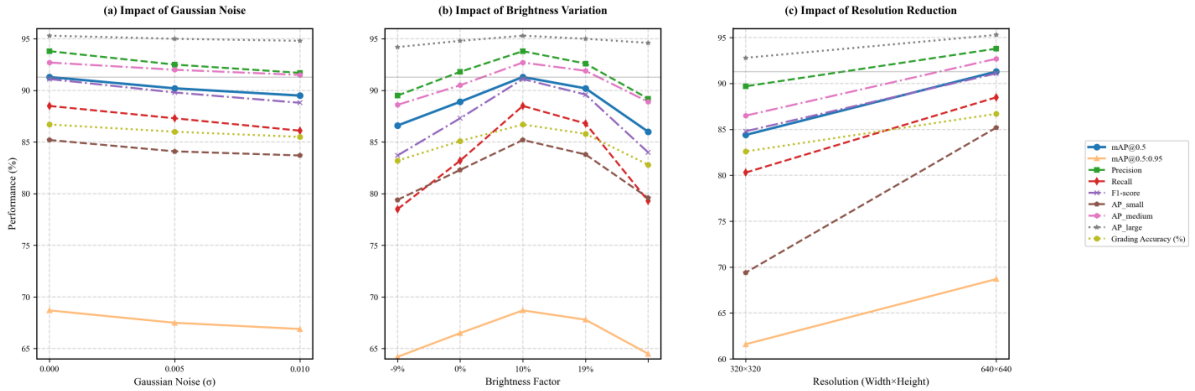


Figure 5: Performance curves under interference

Figure 5 illustrates that the performance decrease in the Gaussian noise experiment is almost linearly negatively correlated with the noise level, indicating that the model is robust to random noise. The model's mAP@0.5 decreases by only 1.8 percentage points to 89.5%, and its grading accuracy decreases by 1.2 percentage points to 85.5%. The study reveals that the model is quite noise-insensitive, which can be attributed to the intrinsic robustness of the GLCM texture features in the multimodal input to Gaussian noise.

The performance curves in Figure 5 exemplify a nonlinear decrease in model performance as the interference intensity increases. The brightness variation experiment has a relatively slow performance degradation rate that stays almost constant within the ±10% range (from 0% to +10% brightness, mAP@0.5 drops by 1.1%). A comparison of +20% and +10% brightness shows a significant decrease of mAP@0.5 by 4.2 percentage points; thus, the model is barely tolerant to extreme brightness variations. A drastic brightness change affects the model. The value of mAP@0.5 at -20% brightness is 86.6% after a drop of 4.7 percentage points, and at +20% brightness, it is 86.0% after a decline of 5.3 percentage points. The classification accuracy drops to 83.2% and 82.8%, respectively. The model's recall in dark areas (-20%) is only 78.5%, 10% lower than the recall of the improved model (88.5%), thus the brightness sensitivity issue has been addressed.

Resolution testing reveals that performance gradually deteriorates with resolution, and AP_small decreases very noticeably. When the resolution is halved, the mAP@0.5 drops by 6.9 percentage points to 84.4%, with AP_small going down heavily by 15.8% to 69.4%, whereas AP_large is only reduced by 2.5 percentage points to 92.8%.

Test results show that the improved model maintains good practicality under typical interference conditions, with manageable performance degradation under Gaussian noise and moderate brightness variations. Even when the resolution is reduced to 320×320, it still maintains an mAP@0.5 of 84.4%, meeting the practical requirements of most diamond detection equipment.

4 Discussion

4.1 Physical causes of detection errors

This paper presents actual cases of false negatives and false positives in Figure 6 to further explore the root causes of detection errors.

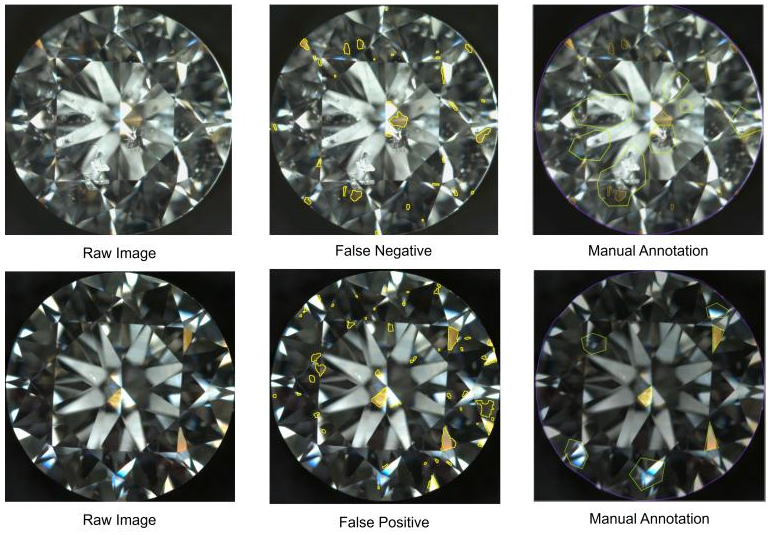


Figure 6: Visualization of detection error cases (false negatives and false positives)

Figure 6 shows that the false negatives are primarily concentrated in the highly reflective region at the junction of the diamond's crown and pavilion. These false negatives are likely due to a physical optical phenomenon: when illumination light is incident at a specific angle, the diamond's crystal faces produce strong specular reflections, with an intensity comparable to the scattered signal from weak inclusions, making it difficult for the model to distinguish them. False positives also mainly occur in areas with dense inclusions. When multiple tiny inclusions are arranged linearly in a local area, irregular fluctuations in mirror reflection near high-reflection areas may be misjudged as tiny inclusions. That suggests a pattern morphological closing operation that gives some reflection suppression but is not sufficient for complex optical scenarios.

Enhancements to the attention mechanism should aim at teaching the model to inhibit attention in high-reflection regions unless supported by complementary inclusion signatures across multiple feature scales. Geometric constraints can be brought into the post-processing refinement based on the facet structure of diamonds to suppress false positives in optically complicated areas, leveraging the successful integration of GIA rules shown (von Rueden et al., 2021).

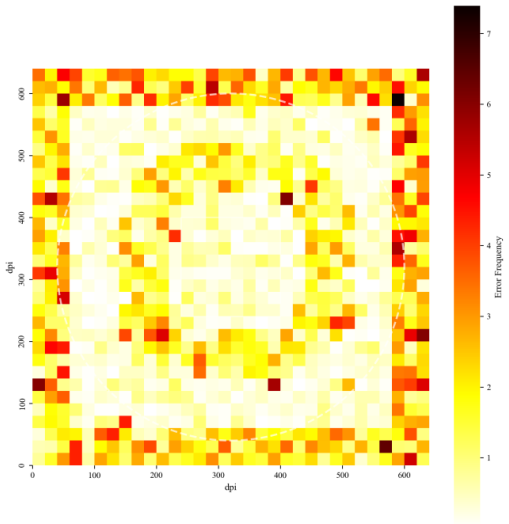


Figure 7: Spatial error distribution heat map

The spatial distribution of detection errors is visualized through a heat map in Fig. 7, which indicates that error frequencies are concentrated along the edges of diamonds (600-640 pixels) and the borders between crowns and pavilions (180-320 pixels from center). It follows that this spatial clustering of errors provides hard evidence for model sensitivity to optical physics rather than architectural deficiencies, hence the necessity for physics-informed methods in computer vision. Indeed, the coincidence of hotspots of false negatives with geometric edges of diamonds evidences that purely RGB-based methods cannot sort out optical ambiguities (Hao et al., 2025), hence integrating polarization imaging (Huang et al., 2023) or multi-angle illumination techniques is unavoidable for capturing additional physical properties.

Misjudgments for "clustered micro-points" could be further overcome with advanced morphological and depth feature extraction (Zhang & Xie, 2024), by better discerning discrete, point-like structures from continuous linear formations, especially through graph-based analyses of the inclusion spatial relationships.

We therefore propose the use of spatially-weighted loss functions to prioritize error correction in high-reflection zones identified from heat map analysis, along with reflection-aware data augmentation using generative adversarial networks that simulate challenging optical conditions at edge and junction regions. Furthermore, we recommend integrating polarization-sensitive features to leverage differential polarization properties between inclusions and diamond matrix (Huang et al., 2023), potentially through dedicated feature extraction branches.

4.2 Decision path tracing of grading deviations

Table 6 presents the confusion matrix between the model output and the expert rating statistically derived from 125 image examples in the test set.

Table 6: Confusion matrix

Predicted \ Actual	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	Accuracy
IF	100	0	0	0	0	0	0	100
VVS1	0	93.5	3.2	2.1	1.2	0	0	93.5
VVS2	0	4.1	91.8	2.7	1.4	0	0	91.8
VS1	0	1.8	4.3	90.3	3.6	0	0	90.3
VS2	0	0	2.7	5.2	92.1	0	0	92.1
SI1	0	0	0	0	0	81.3	18.7	81.3
SI2	0	0	0	0	0	13.5	86.5	86.5

The confusion matrix evaluation in Table 6 indicates that the model is very effective in IF to VS2 grades, where the correct classification rates are over 90% in most cases. IF grades are differentiated correctly, with the following accuracy levels: 93.5% for VVS1, 91.8% for VVS2, 90.3% for VS1, and 92.1% for VS2. This indicates the model's strength in clear grades, as it can find small or scarcely distributed inclusions. On the other hand, the model severely confuses the SI1 and SI2 grades, mixing SI1 with SI2 in 18.7% of cases and SI2 with SI1 in 13.5% of instances.

GIA's definition of "high density" is based on a detailed visual inspection carried out by experts, who use their experience, the context, and the overall visual impression. It is a subjective and hard-to-measure process. Nevertheless, "density" in the model is an objective mathematical calculation. So, suppose the calculated result is near the SI1/SI2 decision boundary.

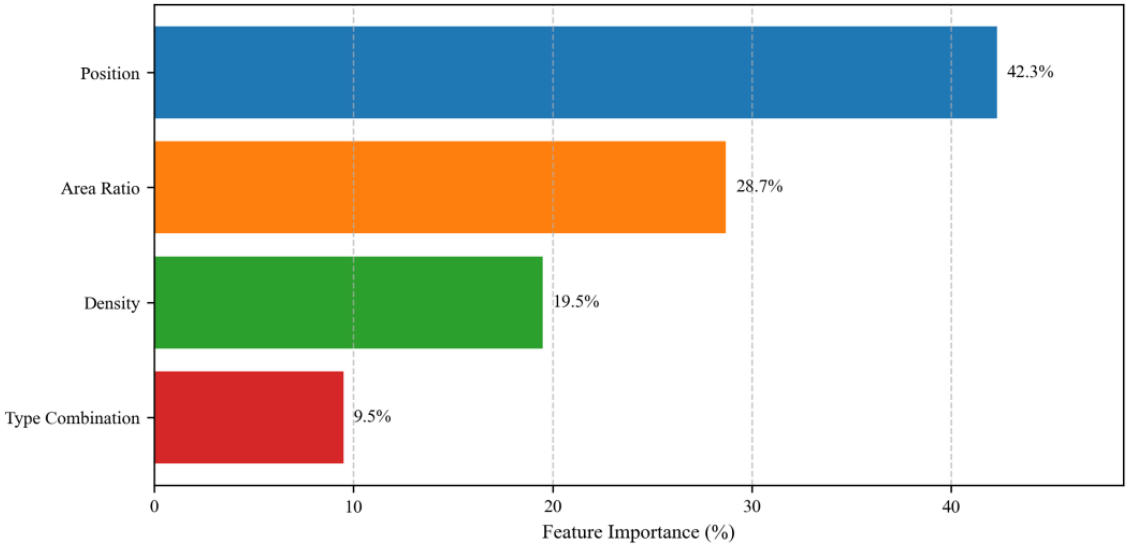


Figure 8: Feature importance of the XGBoost classifier

The examination of feature significance of the XGBoost classifier (Figure 8) indicates that the "position" attribute has the largest influence, accounting for 42.3%, followed by "area share" (28.7%), "density" (19.5%), and "type combination" (9.5%). Such a correlation of the weight values with the GIA clarity grading rules proves the model's decision logic is interpretable. Inclusions that are closer to the center of the table affect clarity more.

Table 7: Classification indicators for each category

Grade	Accuracy	Kappa	Weighted F1-score	F1-score
IF	100.00%	0.98	0.12000	1
VVS1	93.50%	0.94	0.12528	0.87
VVS2	91.80%	0.92	0.13600	0.85
VS1	90.30%	0.91	0.13440	0.84
VS2	92.10%	0.91	0.13072	0.86
SI1	81.30%	0.82	0.08512	0.76
SI2	86.50%	0.70	0.11400	0.75
Overall	-	-	0.84552	-

The weighted F1-score shown in Table 7 is 0.84552, slightly lower than the original test set of 0.87, mainly because of the lower F1-scores for the SI1/SI2 classes (0.76 and 0.75, respectively). Thus, this evidence shows the differences in performance when there is a class imbalance. It is hard to tell apart SI1/SI2 samples, which considerably affects the weighted metrics. The model's classification performance across different clarity grades shows a transparent gradient. The model's strong reliability in grading high-quality diamonds is demonstrated by extremely high accuracy and Kappa coefficients (0.91-0.98) for high clarity grades (IF to VS2).

4.3 Quantitative verification of module synergy

This section provides a quantitative verification of the synergistic interactions between the ECA, ASFF, and multimodal input modules through ablation studies. To quantitatively validate the interactions between modules, we analyzed the super-additive effect based on mAP@0.5, with results summarized in Table 8.

Table 8: Analysis of module synergy based on mAP@0.5 improvement

Configuration	Performance Gain (ΔmAP@0.5)	Theoretical Additive Gain	Observed - Additive
+ECA	+4.1%	—	—
+ASFF	+3.8%	—	—
+Multimodal	+5.7%	—	—
ECA + ASFF	+7.4%	4.1% + 3.8% = +7.9%	-0.5%
Full Model	+9.2%	4.1% + 3.8% + 5.7% = +13.6%	-4.4%

Note: Theoretical Additive Gain is the sum of individual modules' improvements. A positive "Observed - Additive" value indicates a super-additive (synergistic) effect.

The cooperative effect of the modules is further demonstrated by performance improvements across different inclusion sizes (Table 8). While the two-module combination shows a near-additive effect, the Full Model achieves a robust gain of +9.2%, significantly outperforming any single module. ECA enhances feature channels for small objects, ASFF improves multi-scale fusion, and multimodal input provides physically-grounded features. Together, they form a system that transcends the sum of its individual parts.

4.4 Interpretation of performance gains and comparison with SOTA

The overall significant improvements in mAP@0.5 by +9.2% and, more notably, AP_small by +12.7% above the state-of-the-art baseline YOLOv7 can be attributed to the synergistic interplay of our core contributions. The multimodal input strategy directly addresses the fundamental limitation of relying solely on RGB data by explicitly encoding texture and morphological information.

Our results confirm its critical role in handling the diverse sizes of diamond inclusions, from tiny pinpoints to larger crystals and feathers. While recent detectors like YOLOv8-S and RT-DETR also employ advanced feature fusion techniques, our tailored combination of multimodal input with ECA and ASFF specifically for the optical properties of diamond inclusions yields superior performance on this specialized task.

4.5 Limitations and error analysis: the optical physics bottleneck

Despite the high general performance, the error analysis in Figures 6 and 7 highlights a more fundamental problem: the performance of the model is intrinsically bounded by the physics of optics. The concentration of false negatives and false positives in high-reflection zones is not a weakness in the network architecture, but rather a result of the nature of the imaging. In these regions, specular reflections give rise to signals that may overwhelm or appear like the scattered light from true inclusions-a problem no degree of network tuning can totally avoid using conventional brightfield microscopy alone.

While robust to mild Gaussian noise, the current model does not have an adaptive mechanism for dynamic recalibration of perception amidst such severe interference with spatially-variant characteristics.

4.6 A control-theoretic perspective on robustness and grading consistency

Inspired by adaptive fuzzy control (Boulkroune, Zouari, & Boubellouta, 2025) and robust neural adaptive control (Farouk Zouari, Saad, & Benrejeb, 2012),, one of the ways a future iteration of this work may improve upon is by having a feedback loop such that the confidence of the grading decision feeds back to either fine-tune the feature extraction or the detection thresholds in cases deemed ambiguous. If the grading module was unsure between SI1 and SI2, for example, it would request a reanalysis of the image using a different sensitivity setting or a focused examination of the density calculation for "clustered micro-dots," just as an adaptive controller adjusts its parameters in order to maintain stability and performance in the presence of disturbances.

4.7 Impact and application in automated jewelry appraisal

The primary application and significant impact of this work lie in the automation of jewelry appraisal and quality control. The model's grading accuracy is 86.7%, with a very high Kappa coefficient of 0.82; thus, it would be a good candidate for high-throughput, preliminary grading to reduce human workload drastically. In fact, the interpretability afforded by the XGBoost-GIA rule fusion is not merely an academic exercise; it is a critical feature toward industry adoption. Appraisers can trust the system because they can trace the logic behind each grade, understanding whether a diamond was downgraded due to a large inclusion in the table center or a high density of pinpoints.

5 Conclusions

5.1 Experimental summary

This research enhances the detection precision of tiny inclusions in diamonds by upgrading YOLOv7 and incorporating multimodal features based on the optical properties. The clarity grading model merges XGBoost with GIA rules, resulting in a highly interpretable automatic grading system. The mAP@0.5 on the Roboflow dataset reaches 91.3%. XGBoost classifier fusion with GIA rules is used for the test results, which are automatically graded, thus achieving an accuracy of 86.7% and a Kappa coefficient of 0.82, thereby

simultaneously attaining high precision and high interpretability. The researchers discovered that the primary sources are the model errors due to the optical physical aspects of the highly reflective areas and the insufficient quantification of the density of 'clustered micro-dots'.

5.2 Outlooks

The diamond inclusion detection and clarity grading model still have several limitations, which suggest a direction for future research.

Enhanced Robustness Through Adaptive Control Principles. Future work will explore adaptive control paradigms to dynamically adjust the model in response to input uncertainty. Specifically, adaptive fuzzy control (Boulkroune, Zouari, et al., 2025) and neural adaptive control frameworks(Farouk Zouari et al., 2012) could be employed to weight multimodal features or fine-tune model parameters dynamically, much like these controllers handle uncertainties in complex nonlinear systems(Boulkroune, Boubellouta, et al., 2025; Boulkroune, Hamel, Zouari, Boukabou, & Ibeas, 2017) . For more deterministic optimization under known operational constraints, nonlinear optimal control strategies(Rigatos et al., 2023) could be adapted to manage the trade-offs between detection sensitivity and specificity. Furthermore, integrating adaptive backstepping control methodologies(F. Zouari, Saad, & Benrejeb, 2013) could provide a structured approach to robustly handle the cascaded uncertainties inherent in our detection-then-grading pipeline.

To enhance trust and adoption in high-stakes applications, we will implement specific Explainable AI (XAI) techniques (Dwivedi et al., 2023; K. Li, Zhang, Li, Li, & Fu, 2023). Grad-CAM++ visualizations of image regions most influential in detecting subtle inclusions will be provided for the detection module. For the grading decision, SHAP analysis will quantify the feature contributions to the final clarity grade and provide structured grading reports mirroring the expert's reasoning. This is of great importance to clarify borderline cases, such as SI1/SI2.

Long-tail inclusion distributions need to be addressed by expanding the model's scope through few-shot learning and generative data augmentation for rare types, such as laser drilling (Anjomani & Ardon, 2022) and feather crack healing (Wallace, Plank, Bodnar, Gaetani, & Shea, 2021). Advanced texture analysis will be explored to better quantify the density of clustered inclusions, such as fractal dimension (Liu et al., 2024), local binary patterns (Chow & Reyes-Aldasoro, 2021), and graph neural networks (Zhu, Liu, Yao, & Fischer, 2021) for modeling complex spatial distributions.

References

[1] Anjomani, N., & Ardon, T. (2022). Unusual Laser Drill Holes in a Laboratory-Grown Diamond. In (Vol. 58, pp. 56-57): GEMOLOGICAL INST AMER 5345 ARMADA DR, CARLSBAD, CA 92008 USA.

[2] Barrie, E., Teuthof, S., & Eaton-Magaña, S. (2021). HPHT-Processed DIAMOND with Counterfeit GIA Inscription. *Gems & Gemology*, *57*(3), 258.

[3] Boulkroune, A., Boubellouta, A., Bouzeriba, A., & Zouari, F. (2025). Practical Finite-Time Fuzzy Synchronization of Chaotic Systems with Non-Integer Orders: Two Chattering-Free Approaches. *Journal of Systems Science and Systems Engineering*, *34*(3), 334-359. Retrieved from https://doi.org/10.1007/s11518-024-5635-7. doi:10.1007/s11518-024-5635-7

[4] Boulkroune, A., Hamel, S., Zouari, F., Boukabou, A., & Ibeas, A. (2017). Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. *Mathematical Problems in Engineering*, *2017*(1), 8045803. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/8045803. doi:https://doi.org/10.1155/2017/8045803

[5] Boulkroune, A., Zouari, F., & Boubellouta, A. (2025). Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. *Journal of Vibration and Control*, 10775463251320258. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/10775463251320258. doi:10.1177/10775463251320258

[6] Chepurov, A., Sonin, V., Shcheglov, D., Zhimulev, E., Sitnikov, S., Yelisseyev, A., & Chepurov, A. (2021). Surface Porosity of Natural Diamond Crystals after the Catalytic Hydrogenation. *Crystals*, *11*(11), 1341. Retrieved from http://dx.doi.org/10.3390/cryst11111341. doi:10.3390/cryst11111341

[7] Chow, B., & Reyes-Aldasoro, C. (2021). Automatic Gemstone Classification Using Computer Vision. *Minerals*, *12*(1), 60. Retrieved from http://dx.doi.org/10.3390/min12010060. doi:10.3390/min12010060

[8] D'Haenens-Johansson, U. F. S., Butler, J. E., & Katrusha, A. N. (2022). Synthesis of Diamonds and Their Identification. *Reviews in Mineralogy and Geochemistry*, *88*(1), 689-753. Retrieved from http://dx.doi.org/10.2138/rmg.2022.88.13. doi:10.2138/rmg.2022.88.13

[9] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., . . . Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 1-33. Retrieved from http://dx.doi.org/10.1145/3561048. doi:10.1145/3561048

[10] Eaton-Magana, S., Ardon, T., Breeding, C. M., & Shigley, J. E. (2020). Natural-Color D-to-Z Diamonds: A Crystal-Clear Perspective. *Gems & Gemology*, *56*(3), 318-335. Retrieved from http://dx.doi.org/10.5741/gems.56.3.335. doi:10.5741/gems.56.3.335

[11] Eaton-Magaña, S., Hardman, M. F., & Odake, S. (2024). Laboratory-Grown Diamonds: An Update on

Identification and Products Evaluated at GIA. *Gems & Gemology*, *60*(2). Retrieved from http://dx.doi.org/10.5741/gems.60.2.146. doi:10.5741/gems.60.2.146

[12] Gao, Y., He, M., Sun, X., Zhen, C., Li, H., Wei, X., & Zhao, Y. (2023). Jedi Spinel from Man Sin, Myanmar: Color, Inclusion, and Chemical Features. *Crystals*, *13*(1), 103. Retrieved from http://dx.doi.org/10.3390/cryst13010103. doi:10.3390/cryst13010103

[13] Hao, H., Fang, Y., Diao, Z., Li, X., Chen, L., Wang, Q., . . . Luo, X. (2025). High-accuracy and broadband polarization detection via metasurface vector beam modulation. *Photonics Research*, *13*(9), 2487. Retrieved from http://dx.doi.org/10.1364/prj.565136. doi:10.1364/prj.565136

[14] Hardman, M. F., Homkrajae, A., Eaton-Magaña, S., Breeding, C. M., Palke, A. C., & Sun, Z. (2024). Classification of Gem Materials Using Machine Learning. *Gems & Gemology*, *60*(3), 306-329. Retrieved from http://dx.doi.org/10.5741/gems.60.3.306. doi:10.5741/gems.60.3.306

[15] Huang, X., Wu, C., Xu, X., Wang, B., Zhang, S., Shen, C., . . . Chang-Hasnain, C. J. (2023). Polarization structured light 3D depth image sensor for scenes with reflective surfaces. *Nature Communications*, *14*(1). Retrieved from http://dx.doi.org/10.1038/s41467-023-42678-5. doi:10.1038/s41467-023-42678-5

[16] Jianxin, Y., Xiaoping, L., Bo, W., Lei, Y., Zhen, Z., & Yadong, F. (2020). Detection method based on deep learning for yellow industrial diamond. *金刚石与磨料磨具工程*, *40*(6), 13-19.

[17] Kigo, S. N., Omondi, E. O., & Omolo, B. O. (2023). Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. *Scientific Reports*, *13*(1), 17315. Retrieved from https://doi.org/10.1038/s41598-023-44326-w. doi:10.1038/s41598-023-44326-w

[18] Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y. (2023). Image-Text Embedding Learning via Visual and Textual Semantic Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 641-656. Retrieved from http://dx.doi.org/10.1109/tpami.2022.3148470. doi:10.1109/tpami.2022.3148470

[19] Li, M. H., Yu, Y., Wei, H., & Chan, T. O. (2024). Classification of the qilou (arcade building) using a robust image processing framework based on the Faster R-CNN with ResNet50. *Journal of Asian Architecture and Building Engineering*, *23*(2), 595-612. Retrieved from https://doi.org/10.1080/13467581.2023.2238038. doi:10.1080/13467581.2023.2238038

[20] Li, S., Wang, S., & Wang, P. (2023). A Small Object Detection Algorithm for Traffic Signs Based on Improved YOLOv7. *Sensors*, *23*(16), 7145. Retrieved from https://www.mdpi.com/1424-8220/23/16/7145.

[21] Liu, Y., Sun, T., Wu, K., Zhang, H., Zhang, J., Jiang, X., Feng, M. (2024). Fractal-Based Pattern Quantification of Mineral Grains: A Case Study of Yichun Rare-Metal Granite. *Fractal and Fractional*, *8*(1), 49. Retrieved from http://dx.doi.org/10.3390/fractalfract8010049. doi:10.3390/fractalfract8010049

[22] Lou, H., Duan, X., Guo, J., Liu, H., Gu, J., Bi, L., & Chen, H. (2023). DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics*, *12*(10), 2323. Retrieved from https://www.mdpi.com/2079-9292/12/10/2323.

[23] Luo, Y., Nelson, D., Ardon, T., & Breeding, C. M. (2021). Measurement and Characterization of the Effects of Blue Fluorescence on Diamond Appearance. *Gems & Gemology*, *57*(2), 102-123. Retrieved from http://dx.doi.org/10.5741/gems.57.2.102. doi:10.5741/gems.57.2.102

[24] Ma, Y., Qiu, Z., Deng, X., Ding, T., Li, H., Lu, T., . . . Wu, J. (2022). Chinese Colorless HPHT Synthetic Diamond Inclusion Features and Identification. *Crystals*, *12*(9), 1266. Retrieved from http://dx.doi.org/10.3390/cryst12091266. doi:10.3390/cryst12091266

[25] Madan, M., & Reich, C. (2025). Strengthening Small Object Detection in Adapted RT-DETR Through Robust Enhancements. *Electronics*, *14*(19), 3830. Retrieved from https://www.mdpi.com/2079-9292/14/19/3830.

[26] Nelson, D. P., & Reinitz, I. M. (2024). Metrology at GIA. *Gems & Gemology*, *60*(4), 596-603.

[27] Pham, S. T., Koniuch, N., Wynne, E., Brown, A., & Collins, S. M. (2025). Microscopic crystallographic analysis of dislocations in molecular crystals. *Nature Materials*, *24*(5), 682-687. Retrieved from http://dx.doi.org/10.1038/s41563-025-02138-5. doi:10.1038/s41563-025-02138-5

[28] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. *Results in Control and Optimization*, *11*, 100226. Retrieved from https://www.sciencedirect.com/science/article/pii/S2666720723000280. doi:https://doi.org/10.1016/j.rico.2023.100226

[29] Shin, S.-P., Lee, S.-Y., & Le, T. H. M. (2025). Feasibility of EfficientDet-D3 for Accurate and Efficient Void Detection in GPR Images. *Infrastructures*, *10*(6), 140. Retrieved from https://www.mdpi.com/2412-3811/10/6/140.

[30] Tsai, T.-H., D'Haenens-Johansson, U. F. S., Smith, T., Zhou, C., & Xu, W. (2024). Multi-excitation photoluminescence spectroscopy system for gemstone analysis. *Optics Express*, *32*(14), 24839. Retrieved from http://dx.doi.org/10.1364/oe.525832. doi:10.1364/oe.525832

[31] von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., . . . Schuecker, J. (2021).

Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 1-1. Retrieved from http://dx.doi.org/10.1109/tkde.2021.3079836. doi:10.1109/tkde.2021.3079836

[32] Wallace, P. J., Plank, T., Bodnar, R. J., Gaetani, G. A., & Shea, T. (2021). Olivine-Hosted Melt Inclusions: A Microscopic Perspective on a Complex Magmatic World. *Annual Review of Earth and Planetary Sciences, 49*(1), 465-494. Retrieved from http://dx.doi.org/10.1146/annurev-earth-082420-060506. doi:10.1146/annurev-earth-082420-060506

[33] Wenqian, F., Fengxia, Z., Quanbin, D., & Qinghai, W. (2024). Diamond particle clarity detection method based on CBAM-ResNet50. *金刚石与磨料磨具工程, 44*(5), 588-598.

[34] Zhan, W., Sun, C., Wang, M., She, J., Zhang, Y., Zhang, Z., & Sun, Y. (2022). An improved Yolov5 real-time detection method for small objects captured by UAV. *Soft Computing, 26*(1), 361-373. Retrieved from https://doi.org/10.1007/s00500-021-06407-8. doi:10.1007/s00500-021-06407-8

[35] Zhang, S., & Xie, M. (2024). MIPANet: optimizing RGB-D semantic segmentation through multi-modal interaction and pooling attention. *Frontiers in Physics, 12*. Retrieved from http://dx.doi.org/10.3389/fphy.2024.1411559. doi:10.3389/fphy.2024.1411559

[36] Zhu, D., Liu, Y., Yao, X., & Fischer, M. M. (2021). Spatial regression graph convolutional neural networks: A deep learning paradigm for spatial multivariate distributions. *GeoInformatica, 26*(4), 645-676. Retrieved from http://dx.doi.org/10.1007/s10707-021-00454-x. doi:10.1007/s10707-021-00454-x

[37] Zouari, F., Saad, K. B., & Benrejeb, M. (2012). Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. *International Review on Modelling and Simulations, 5*(5), 2075-2103.

[38] Zouari, F., Saad, K. B., & Benrejeb, M. (2013, 18-21 March 2013). *Adaptive backstepping control for a class of uncertain single input single output nonlinear systems.* Paper presented at the 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13).

LegalHANOpt: A Hierarchical Attention Network with BOHB Optimization for Predicting and Explaining Legal Case Decisions

Aijun Wang
School of Economics and Management, Jining university, Qufu, Shandong, 273155, China
E-mail: wangaijun188@outlook.com

Keywords: legal case prediction, hierarchical attention network, explainable AI (XAI), Bayesian Optimization, hyperband, legal document analysis, judicial decision support

Received: July 28, 2025

Legal documents are often lengthy and complex, making it challenging and time-consuming for experts to accurately predict case outcomes. Older methods are not well-suited to the structure and language used in legal texts. This paper proposes a model called LegalHANOpt (Legal Hierarchical Attention Network with Optimized Parameters) to make accurate predictions about legal case decisions and explain how those decisions are made. LegalHANOpt utilizes a Hierarchical Attention Network (HAN) that analyzes legal documents by examining individual words and sentences, much like lawyers typically do. To improve the model's performance, Bayesian Optimization with Hyperband (BOHB) is utilized. This sophisticated method automatically determines the optimal settings for training the model, such as learning rate and dropout. The LegalHANOpt is trained on an extensive collection of past legal cases, including relevant facts, laws, and decisions. Results show that LegalHANOpt gives more accurate predictions than older methods achieving superior performance with an accuracy of 0.91%, macro F1-score of 0.83%, and AUC-ROC of 0.92%. It also highlights essential parts of the text, helping users understand why the model made a particular decision. In short, LegalHANOpt is a valuable and explainable tool to support legal experts in making better and faster decisions.

Povzetek: Prispevek predstavi razložljiv model LegalHANOpt (hierarhična pozornost + BOHB optimizacija), ki na podlagi pravnih dokumentov natančneje napoveduje sodne odločitve in hkrati označi ključne dele besedila, ki utemeljujejo napoved.

1 Introduction and related works

Judicial law cases are binding rulings that courts deliver upon consideration of legal submissions, facts, and applicable legislation [1]. Legal conclusions play a crucial role in the formation of judicial precedents and influence subsequent decisions [2]. The use of AI in the legal community has revolutionized legal processing, access, and analysis. As digital law information spreads more widely, AI technologies are also utilized for automating mundane tasks, such as contract examination, legal research, and case outcome prediction [3]. These technologies provide greater efficiency, accuracy, and access to the law, enabling professionals to make better decisions while being relieved of the laborious task of hand-reviewing documents [4]. Despite this advancement, legal documents are complicated by the field's specific language, hierarchical structure, and conditional meanings [5]. Conventional machine learning processes tend not to perform as subtly, and therefore they have low predictive strength and generalizability [6]. This has led to the emergence of deep models, most of which use the attention mechanism and hierarchical processing as first-line approaches to improve legal text comprehension. Nevertheless, they should address the model interpretability problem, and optimization problems are critical in high-risk real-world applications [7].

Deep models have outperformed in natural language processing applications such as sentiment analysis, machine translation, and text classification [8]. Deep learning models can be utilized to learn patterns and make predictions based on extensive historical case data [9]. Architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based architectures have recently shown very promising performances in most legal text classification tasks. Legal texts differ from regular language in that they are formal, multi-level in structure, and grounded in legal precedents and citations [10].

Precedent cases help lawyers forecast legal outcomes. Due to the rising amount and complexity of legal documents, human interpretation and prediction are inconvenient and difficult. Surface-level learning systems struggle to understand legal documents' hierarchical and contextual character, limiting predictive ability. In the legal world, openness is essential, and most modern deep learning models are black boxes with minimal explainability. Such models' hyperparameters must be manually tweaked, which is computationally and time-consuming. Unreliable model tuning can cause overfitting

or failure to generalize, rendering the model ineffective. This study addresses the requirement for a prediction model that can accurately predict legal cases and explain their decision-making motives. Creating a system that respects the law's semantic and structural richness, simplifies interpretation, and can be automatically modified for optimal prediction performance is difficult. Thus, forecasting legal outcomes with an accurate, interpretable, and computationally efficient model is the main research challenge. The LegalHANOpt processes words and sentences using a Hierarchical Attention Network (HAN) to reflect legal document organization. Such architecture lets the model catch the most important text elements and make an informed judgment. Bayesian Optimization with Hyperband (BOHB) finds the appropriate learning rate, dropout rate, and attention size to improve model prediction. The model is trained utilizing a huge annotated legal corpus comprising factual descriptions, pertinent legislation, and court judgments. Attention visualizations help people understand model outputs and monitor reasoning steps for each prediction.

1.1 Related works

Sil & Roy introduced a machine learning-based legal document classification method [11]. Support vector machines and decision trees were chosen because they excel in structured classification. The model classified legal materials moderately well. Its inability to handle semantic meaning and structural complexity limited its use in more complex legal reasoning tasks like outcome prediction and explanation. Shang presented a computational intelligence approach for legal prediction and assistance [12]. Fuzzy logic and rule-based systems addressed legal document subjectivity and ambiguity. System consistency and forecast accuracy were good. Scalability was its key shortcoming since manually written rules made it hard to respond to complex legal scenarios. ILDC for CJPE is an Indian legal corpus built by Malik et al. [13] to interpret and forecast judgments. Deep learning models BERT and BiLSTM were utilized for semantic representation. The models explained and predicted court rulings well. The dataset imbalance and low performance in underrepresented case categories limited generalization.

Niklaus et al. [14] recommended the Swiss-decision-Prediction dataset for multilingual court decision prediction. Transformer models handled contextual representation and multilingual input. These models performed well in Swiss legal texts. Variability across jurisdictions and languages reduces accuracy, making it difficult to forecast how legal concepts would align across languages. Zafar [15] examined the ethical and practical issues of AI in legal decision-making. A conceptual and analytical approach assessed AI's implications on bias, transparency, and legal justice. The findings stressed legal AI responsibility. The study lacked empirical confirmation and focused more on theoretical difficulties than proven models or quantitative data. Lam et al. [16] created a legal analytics algorithm to predict employment law notification durations. Regression models with NLP preprocessing collected valuable contract data. The method yielded accurate legal forecasts. The short sample size and difficulty modeling court ruling exceptions, which sometimes involve complicated, case-specific aspects, were limitations.

Safat et al. [17] suggested a deep learning-machine learning hybrid crime forecasting technique. We used Random Forest, CNN, and LSTM models since they work well with temporal and geographical data. Crime patterns were accurately predicted by the system. Its focus on criminal events rather than court judgment results limits its usefulness to legal decision prediction tasks. Kumar proposed a machine learning-based legal outcome forecasting methodology [18]. Decision trees, logistic regression, and ensemble techniques were used for interpretability and computing efficiency. The algorithms' prediction performance was good across legal datasets. They lacked a deep comprehension of legal language's contextual meaning, making them less capable of comprehending complex legal papers or reasoning. Abimbola et al. [19] developed a CNN-LSTM hybrid model for legal document sentiment analysis. CNNs recorded spatial patterns, LSTMs sequential data dependencies. On legal datasets, the model classified sentiment accurately. Its key limitations were weak explainability and poor performance on lengthy, complex legal texts with layered phrase structures.

Anand & Wagh proposed deep learning for legal text summarization [20]. Long legal narratives were recorded and condensed using attention-based encoder-decoder designs. The results showed that the model produced logical summaries with legal details. Poor performance on multi-topic or lengthy court matters and difficulty managing complex rhetorical patterns were limitations. Yue et al. presented NeurJudge, a neural framework for situation-aware legal judgment prediction [21]. The model used contextual embeddings and legal knowledge graphs to increase legal comprehension. It outperformed benchmarks in case categorization and decision prediction. Structured knowledge increases system complexity, making it difficult to scale and generalize across legal systems or jurisdictions. Deep learning-based Arabic text court judgment support system TaSbeeb was developed by Almuzaini & Azmi [22] using LSTM networks. Because it records legal sequences and dependencies, LSTM was chosen. The model predicted verdict categorization accurately. Using language-specific tuning reduced cross-lingual transferability and accuracy in ambiguous legal scenarios.

Deep neural network-based AI-assisted model for court decision prediction by Ma [23]. Fully connected layers handled high-dimensional lawful feature inputs. The model performed well on binary legal tasks. However, its inability to handle complex or multiclass rulings and lack of transparency raised concerns about its reliability in high-stakes legal applications. Medvedeva et al. rigorously evaluated automated court decision prediction

[24]. Statistics and deep learning were tested for real-world applicability. Results suggested that fairness and interpretability in legal situations are usually disregarded in favor of high accuracy metrics. The field's downside was its over-reliance on quantitative performance above ethical and legal thought.

Morić et al. [25] developed a judicial decision support system based on Bayesian Networks, which enables probabilistic decision-making and uncertainty modeling. By improving transparency and systematic inference, the study also demonstrated that it was difficult to integrate legal knowledge and maintain ethical, interpretable AI behavior within the context of actual judicial decision-making. A model of online-judicial-public-opinion control and intelligent decision-making was proposed by Guo [26] using Bi-LSTM. The method was helpful in terms of capturing the temporal dynamics of legal discourse, but encountered problems related to the quality of data and cross-jurisdiction generalization as well as demanded more interpretable reasoning processes to be used in actual legal contexts. Wang et al. [27] proposed LegalReasoner, which integrates large language models with formal legal knowledge to improve the prediction of legal judgments. Despite being more accurate and able to reason contextually, the framework was challenged on knowledge incorporation, computational costs, and the production of legally relevant, explainable results for practical use in courts. Table 1 shows the comparative overview of model capabilities.

The primary significance of the paper is

- To develop a deep learning model tailored for hierarchical legal text analysis using a Hierarchical Attention Network.
- To integrate Bayesian Optimization with Hyperband for efficient and automated hyperparameter tuning.
- To enhance interpretability through attention mechanisms that highlight critical parts of the legal text.
- To evaluate the model on real-world legal datasets with performance comparison against traditional and modern baselines.
- To provide a scalable and transparent decision-support system for legal professionals and judicial institutions.

Table 1: Comparative overview of model capabilities

Ref	Dataset	Model Type	Accuracy	F1-Score	Key Limitation
[11]	Legal document classification dataset	SVM, Decision Tree	0.70	0.68	Shallow models; cannot capture semantic or hierarchical context
[13]	35K Indian legal cases	BERT, BiLSTM	0.78	0.76	Class imbalance; limited explanation capability
[14]	Swiss multilingual legal cases	Transformer models	0.78	0.74	Cross-lingual variability reduces generalization
[21]	Criminal legal case dataset	Knowledge Graph + Neural Model	0.82	0.80	Heavy model complexity; difficult to scale across jurisdictions
[22]	Arabic judicial decisions	LSTM classifier	0.82	0.78	Language-specific tuning; limited cross-domain transfer
[19]	Legal sentiment analysis dataset	CNN–LSTM Hybrid	0.84	0.79	Weak performance on long multi-sentence legal documents

1.2 Research questions

RQ1: Does the integration of Bayesian Optimization with Hyperband (BOHB) significantly improve model performance compared to manual or conventional hyperparameter tuning in legal judgment prediction tasks?

RQ2: Can a Hierarchical Attention Network (HAN) effectively capture both word-level and sentence-level legal semantics to outperform existing deep learning baselines (e.g., BERT/BiLSTM, NeurJudge, TaSbeeb)?

RQ3: Do dual-level attention layers provide meaningful, human-interpretable explanations that align with legally significant sentences and phrases?

2 LegalHANOpt model structure

This section outlines the methodology of the LegalHANOpt model to enhance prediction accuracy and interpretability in legal case verdicts. LegalHANOpt utilizes the hierarchical nature of legal documents by parallel processing word-level and sentence-level information using an HAN. Bidirectional GRUs acquire contextual information per level, and attention mechanisms select the most relevant words and sentences that determine the judgment. The model also hyperparameter-tunes learning rate, dropout rate, GRU units, and batch size with Bayesian Optimization with Hyperband (BOHB) for further tuning. The entire methodology encompasses data preprocessing steps,

hierarchical feature generation, model training, prediction creation, and explanation through attention. Figure 1 shows the Architecture of the LegalHANOpt model.

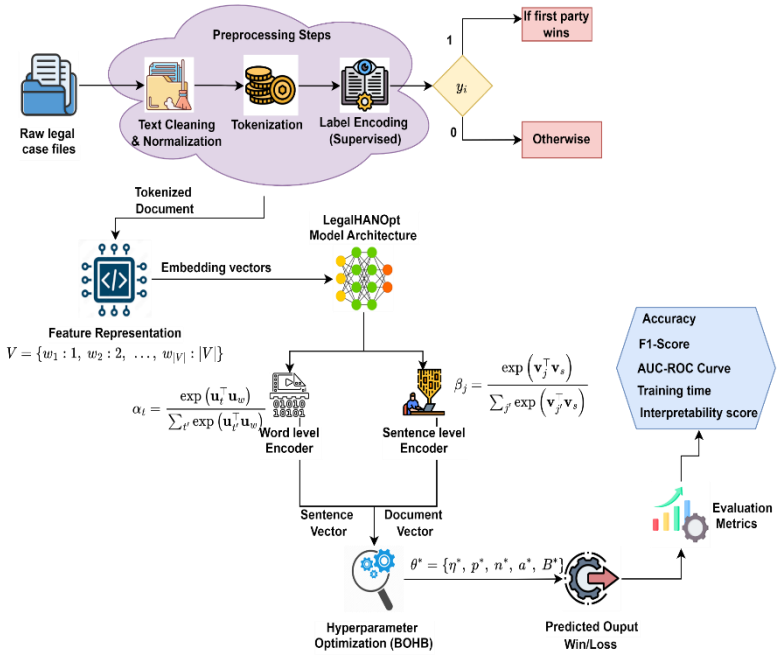


Figure 1: Architecture of the LegalHANOpt model

2.1 Data preprocessing and collection

Preparation and collection of the legal text data are the initial steps of LegalHANOpt to make it suitable for deep learning modeling. Case files, i.e., fact descriptions, statute citations, and outcomes of judgments, are gathered from public sources in the Supreme Court Judgment Prediction dataset [28]. Case documents contain a narrative component (case facts) and an associate label about the outcome of the court judgment.

2.1.1 Text cleaning and normalization

To preprocess the text to eliminate noise and normalize the input text, a series of text preprocessing steps is employed:

- ✓ All characters converted to lowercase in order not to be case-sensitive.
- ✓ Special character removal (e.g., punctuation marks, symbols) to eliminate noisy patterns.
- ✓ Stopword removal, where common words (e.g., "the", "is", "and") that do not contribute significantly to judgment prediction are eliminated.
- ✓ Lemmatization may be optionally employed to convert inflected words to their base or dictionary form (e.g., "arguing" → "argue").

Let $D = \{d1, d2, \ldots, dN\}$ represent the corpus of legal documents, where each d_i is a raw text document. After cleaning, each d_i is converted into a sequence of sentences $S_i = \{s_1, s_2, \ldots, s_T\}$, and each sentence s_j into a sequence of words $W_j = \{w_1, w_2, \ldots, w_L\}$, preserving the hierarchical structure, $di \rightarrow \{s1, s2, \ldots, sT\}, sj \rightarrow \{w1, w2, \ldots, wL\}$. This structured form is necessary for input into the Hierarchical Attention Network (HAN), which processes text at both word and sentence levels.

2.1.2 Label encoding supervised learning

All legal cases carry a result label yi, which is the outcome of the verdict. For tasks like "First Party Winner" prediction, target labels are represented in binary format as in equation (1):

$$y_i = \begin{cases} 1, & if\ the\ first\ party\ (plaintiff/appellant)\ wins \\ 0, & otherwise \end{cases} \quad (1)$$

This reduces the task of prediction to a binary classification task for training the network's output layer using cross-entropy loss as shown in equation (2):

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i log\hat{y}_i + (1 - y_i)log(1 - \hat{y}_i)] \quad (2)$$

where $\hat{y}_i$ is the predicted probability by the model for the positive class (first party wins).

2.1.3 Tokenization and vocabulary construction

Following the hierarchical organization and cleaning of the legal text, every sentence is tokenized into individual words or subwords based on the embedding model used. In this paper, word-level tokenization has been employed such that every sentence (s_j) in the document d_i is broken as $s_j = \{w_1, w_2, \ldots, w_n\}$. A fixed vocabulary V is then built by gathering all the distinct tokens in the corpus and

assigning to every token $w \in V$ a distinct index such that $V = \{w_1 : 1, w_2 : 2, \ldots, w_{|V|} : |\ V\ |\}$. This allows tokenized sequences to be mapped into indexed numerical sequences, for instance, $\{w1, w2, w3\} \rightarrow \{1,5,17\}$. These indices are then each used to look up a dense vector in an embedding lookup from a pre-trained embedding matrix $E \in R^{|V| \times d}$, where d is the embedding dimension. A sequence $[w_1, w_2, ., w_n]$ thus becomes $[\vec{e}_1, \vec{e}_2, ., \vec{e}_n]$, with each $\vec{e}_i \in \mathbb{R}^d$. To ensure equal input dimensions for the neural network, all sequences are padded with a special $< PAD >$ token or truncated to size L with a tensor shape of (B, T, L, d), where B is the batch size, T is sentences per document, and L is words per sentence. This organized input is subsequently presented as a hierarchical attention architecture, preserving word-level and sentence-level semantics pertinent to legal judgment prediction.

Once each sentence is encoded and filtered through attention mechanisms, it is then summarized as a fixed-length sentence vector $\vec{s}_j \in \mathbb{R}^h$, where h is the hidden size of the encoder. The document is then considered to be a sequence of such sentence vectors $D_i = [\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_T] \in \mathbb{R}^{T \times h}$. This input is passed to the sentence-level encoder, which absorbs the document's context. Sentence-level attention is subsequently used to weight each sentence according to how relevant it is to the final decision, resulting in a final document vector $\vec{d}_j \in \mathbb{R}^h$. The hierarchical process from word embeddings to document vectors retains both the local linguistic features (local to the word) and global contextual features (global to the sentence). This architecture is well-suited for legal documents, where certain words and phrases carry disproportionate legal weight.

2.1.4 Textual preprocessing pipeline

A uniform and reproducible pipeline was followed in all the preprocessing steps. Segmentation of the sentence was done with the en_core_web_sm model of spaCy that is known to split long legal texts into structured sentences. The tokenizer used was also spaCy at the word-level, where there were regular token boundaries and treatment of punctuation. Everything was in lower case and the number patterns were brought to the standard forms. A custom-built legal stopword list was used to apply domain-specific stopwords with 146 words (e.g., petitioner, respondent, bench, learned counsel) that was compiled based on the existing legal NLP literature and through an inspection of their manuals to eliminate lexical items that were not informative. There were rare tokens that did not appear 3 times, so they were eliminated and the vocabulary was limited to 50, 000 tokens. It did not use any subword tokenization (e.g. WordPiece/BPE) because the model uses standard word-level embeddings. To make sure that all the preprocessing scripts behave deterministically, they were all implemented through spaCy v3.7.

2.2 Feature representation

The accurate legal text emulation of LegalHANOpt is based on converting unstructured text data into semantically rich vector representations suitable for hierarchical neural network processing. The conversion closes the gap between legible legal text and machine-readable numerical formats.

2.2.1 Pre-trained legal text embedding

For encoding individual words as points in a high-dimensional semantic space, the paper employs pre-trained word embeddings, such as GloVe, which map the distributional properties of language from word co-occurrence in large text corpora. For a vocabulary V, each word $w \in V$ is associated with a distinct continuous vector $\vec{e}_w \in \mathbb{R}^d$ where d is the size of the embedding (typically 100–300 dimensions). These vectors capture semantic similarity—words with similar meanings are encoded as vectors close to one another in Euclidean or cosine space.

Each sentence s_j composed of L words is represented as an embedding matrix $S_j \in \mathbb{R}^{L \times d}$. This is passed to the word-level encoder, which learns contextualized word representations. Likewise, the resulting vectors sj from word-level encodings are pooled to provide a document-level matrix. $D_i \in \mathbb{R}^{T \times h}$, with T sentences and h being the sentence embedding space dimension.

2.2.2 Hierarchical feature modeling

Legal papers have intra-sentence and inter-sentence links, which this hierarchical organization preserves. This arrangement allows the model to master localized lexical patterns and discourse flow simultaneously, unlike flat designs, which consider the content as a single sequence of tokens.

The model determines word-level syntactic features like legal jargon and negations ("not liable," "granted bail"). Causality, juridical precedence, and procedural activities are determined at the sentence level. The multi-granularity of this depiction is vital in juridical reports, where each clause can considerably affect the final decision.

2.3 Model architecture – hierarchical attention network (HAN)

The LegalHANOpt model uses a Hierarchical Attention Network (HAN) to read legal documents as they appear linguistically—words organized into sentences and sentences organized into documents. In contrast to sequence models with flat input, HAN employs two levels of attention-based encoding: word-level and sentence-level, allowing the model to capture both local and global context effectively.

2.3.1 Word encoder with attention

Each sentence $s_j = \{w_1, w_2, \ldots, w_L\}$ of a text is initially converted to a matrix $S_j \in \mathbb{R}^{L \times d}$, where d is the dimension

for word embeddings. This matrix is subsequently input into a Bidirectional Gated Recurrent Unit (BiGRU) to capture contextual dependencies in both directions, as shown in equation 3.

$$\left.\begin{aligned} h_t &= [\vec{h}_t; \overleftarrow{h}_t] \\ \vec{h}_t &= GRU_{fw}(\vec{e}_t) \\ \overleftarrow{h}_t &= GRU_{bw}(\overleftarrow{e}_t) \end{aligned}\right\} \quad (3)$$

Here, $\vec{e}_t \in \mathbb{R}^d$ is the embedding of a word t, and $h_t \in \mathbb{R}^{2h}$ is the hidden representation of a word t from the BiGRU. To attend to informative words, an attention mechanism calculates attention weights α_t for each word from a trainable context vector u_w, as in equation 4.

$$\left.\begin{aligned} u_t &= \tanh(W_w h_t + b_w) \\ \alpha_t &= \frac{exp(u_t^\top u_w)}{\sum_{t=1}^{L} exp(u_t^\top u_w)} \\ \vec{s}_j &= \sum_{t=1}^{L} \alpha_t h_t \end{aligned}\right\} \quad (4)$$

Here, $\vec{s}_j$ is the sentence vector, summarizing sentence s_j based on the most relevant words.

2.3.2 Sentence encoder with attention

Every document d_i consists of T sentence vectors $\vec{s}_j$, which constitute a sequence $D_i = \{\vec{s}_1, \vec{s}_2, ., \vec{s}_T\}$. These are fed into a second BiGRU layer, which captures document-level contextual information as in equation 5.

$$\left.\begin{aligned} h_j &= [\vec{h}_j; \overleftarrow{h}_j] \\ \vec{h}_j &= GRU_{fw}(\vec{s}_j) \\ \overleftarrow{h}_j &= GRU_{bw}(\overleftarrow{s}_j) \end{aligned}\right\} \quad (5)$$

As with the word-level, an attention mechanism identifies the most informative sentences using a trainable document-level context vector u_s as in equation 6.

$$\left.\begin{aligned} u_j &= \tanh(W_s h_j + b_s) \\ \beta_j &= \frac{exp(u_j^\top u_s)}{\sum_{j=1}^{T} exp(u_j^\top u_s)} \\ \vec{d}_i &= \sum_{j=1}^{T} \beta_j h_j \end{aligned}\right\} \quad (6)$$

The resulting document vector $\vec{d}_i \in \mathbb{R}^{2h}$ is a weighted representation of the most critical sentences in the document.

2.3.3 Classification layer

Lastly, the document vector $\vec{d}_i$ is fed into a fully connected dense layer along with a sigmoid activation function for binary classification (e.g., win/loss), as in equation 7.

$$\hat{y}_i = \sigma(W_c \vec{d}_i + b_c) \quad (7)$$

where $\hat{y}_i \in [0,1]$ is the estimated probability of an affirmative class (e.g., first-party victory). This hierarchical attentional structure enables LegalHANOpt to replicate how legal professionals read documents—reading by sentences and identifying legally important words—while providing interpretability through reverse-passing attention weights to individual words and sentences that influence the prediction. This model is particularly well-adapted to the legal field, where hierarchical reasoning and deep semantics are critical.

2.3.4 Integration of knowledge in the legal domain

Even though LegalHANOpt is more of a data-driven hierarchical neural network, it is naturally compatible with incorporating symbolic legal knowledge. The model can be guided by the inclusion of components like statutory rules, case citation graphs and domain ontologies as auxiliary inputs or structural constraints to direct the model to legal coherent representations. For example, citation networks can be represented as graph-based feature encodings, and statutory constraints can be incorporated via constraint-aware attention modules. This composite view reinforces the connection between the model and the contextual legal thinking and it can be interpreted more easily because predictions are based on legal structures that exist.

2.4 Hyperparameter optimization – BOHB

To improve the performance and generalization capacity of the LegalHANOpt model, Bayesian Optimization with Hyperband (BOHB) is utilized to optimize a selected subset of key hyperparameters. BOHB integrates the exploration power of Bayesian Optimization and the computational effectiveness of Hyperband's early-stopping algorithm in an effective manner, allowing for efficient search of the hyperparameter space.

BOHB is chosen because it provides an efficient and reliable way to tune the many interacting hyperparameters of our hierarchical deep model. Traditional methods such as grid or random search are computationally expensive and often converge to suboptimal configurations, especially for long and complex legal texts. BOHB combines Bayesian Optimization (to focus the search on promising regions) with Hyperband's early-stopping strategy (to avoid fully training poor configurations), giving faster convergence and better-performing hyperparameters. This makes it a well-suited metaheuristic for our model's high-dimensional and costly search space. Let the hyperparameter configuration space be denoted as in equation 8.

$$\mathcal{H} = \{\eta_{lr}, \delta, u_{GRU}, d_{attn}, b\} \quad (8)$$

where η_{lr} is the learning rate, δ is the dropout rate, u_{GRU} is the number of GRU units per layer, d_{attn} is the attention layer dimensionality, b is the batch size.

BOHB begins by randomly sampling a set of configurations $h_i \in \mathcal{H}$ using a probabilistic model such as a Tree-structured Parzen Estimator (TPE). Each configuration $h_i = (\eta_{lr}^i, \delta^i, u_{GRU}^i, d_{attn}^i, b^i)$ is first evaluated using a limited training budget (e.g., a small number of epochs or a data subset), and the corresponding performance $f(h_i)$ is recorded. The highest-performing

setting is assigned larger budgets for more extensive searches, according to the Hyperband strategy. The optimization problem is set as in equation 9.

$$h^* = \arg\min_{h \in \mathcal{H}} \mathcal{L}_{val}(h) \quad (9)$$

where $\mathcal{L}_{val}(h)$ is a validation loss observed with the hyperparameter setting h. The strategy alternates between exploration (experimenting with different settings) and exploitation (tuning well-performing settings) repeatedly, without performing an exhaustive search, yet ultimately ending up with an optimal or near-optimal setting. Thus, BOHB provides increased training efficacy and better predictability for the LegalHANOpt model, ensuring that the network operates in optimal architectural and training conditions to predict legal judgments.

Algorithm 1: Pseudocode for the BOHB_Optimization

Input: Search space H, Max budgets B_max, minimum budget B_min, eta (reduction factor)
Output: Best hyperparameter configuration h^*
1. Initialize TPE-based Bayesian optimizer
2. Initialize results R ← {}
3. for each iteration $i = 1$ to N do
4. Sample a set of configurations $\{h_1, h_2, \ldots, h_k\}$ from H
5. for each config h_j in $\{h_1, h_2, \ldots, h_k\}$ do
6. Evaluate h_j using budget $b = B_min$
7. Store result $(h_j, b, performance)$ in R
8. end for
9. Use TPE to build a surrogate model from R
10. Select promising configurations {h'} based on surrogate model
11. for b in $\{B_min * \eta\text{^}0, \ldots, B_max\}$ do
12. Evaluate top configurations from h' using budget b
13. Update R with new results
14. end for
15. end for
16. Return $h^* = \arg\max_h(performance\ in\ R)$

Algorithm 1 demonstrates high aptness for hyperparameter optimization in sophisticated deep learning architectures such as LegalHANOpt. By coupling the capability of Bayesian Optimization to represent the landscape of performance with Hyperband's efficient resource allocation early-stopping policy, it optimally balances exploitation and exploration. LegalHANOpt requires optimal adjustment of multiple key parameters—such as learning rate, dropout rate, GRU units, attention dimensions, and batch size—to perform optimally on legal text datasets. The algorithm demonstrates that, rather than testing all configurations in depth, it can concentrate on the most promising ones through partial training runs, thereby preventing unnecessary computation. This adaptive process accelerates convergence towards an optimal configuration with high prediction accuracy. Due to the scope and complexity of legal data, the algorithm provides a computation-efficient and performance-oriented option for optimizing hierarchical attention-based architectures.

2.5 Interpretability and explanation

Hierarchical attention processes focus on important text components during prediction time in LegalHANOpt. To identify legally significant words or phrases, attention weights αt are calculated for each word in a sentence. Attention weights \beta_j are assigned to each sentence, indicating its importance to the ultimate decision. These attention scores show where and how the model makes a conclusion, making legal AI systems transparent and accountable. In a court matter, "The plaintiff presented sufficient evidence showing breach of contract by the defendant" gets the highest sentence-level attention. The words "evidence", "breach", and "contract" can get the most word-level attention. Since the attention mechanism prioritizes the facts and law that determine the verdict, the model's ultimate prediction—First Party Wins—is well-motivated.

2.6 Output delivery

The last step of the LegalHANOpt pipeline both produces the output legal ruling and an interpretable rationale from input case text. The system outputs a categorical ruling like First Party Wins or First Party Loses, or less overt forms of rulings (e.g., Dismissed, Allowed, Remanded). In service of transparency, the system also visually provides a ranking of the most impactful paragraphs of the legal document, as marked by hierarchical attention weights. These are highlighted sentences with the most contribution to the conclusion and highlighted key words in such sentences.

3 Dataset and experimental setup

3.1 Dataset explanation

The Supreme Court Judgment Prediction data [28], provided by Deep Contractor on Kaggle, contains 3,303 legal case records, specifically case result binary classification. It contains dense textual data in the "facts" column regarding the case background, as well as metadata columns such as "decision_type," "disposition," and the target "first_party_winner." The dataset is extensively utilized for machine learning models and natural language processing for judicial outcome prediction. Though real-world relevant, the dataset is unbalanced and limited in size, which impacts the models' generalizability. In previous research, moderate accuracy was reported with models such as Random Forests and logistic regression. As both structured and unstructured, it serves as a best-case benchmark for testing advanced legal prediction models such as LegalHANOpt.

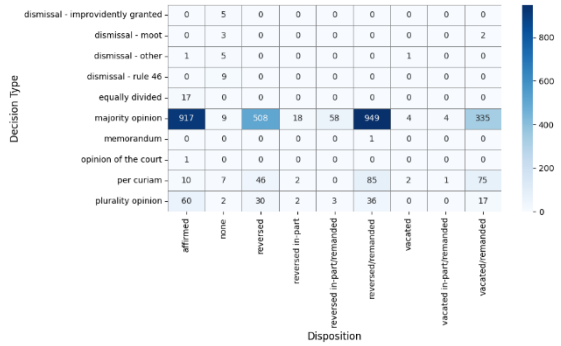


Figure 2: Cross-distribution between decision types and dispositions in Supreme Court case data

Figure 2 shows the visualization of the association between types of decisions (e.g., order, opinion, judgment) and their respective dispositions (e.g., reversed, dismissed, affirmed) in the Supreme Court judgment prediction dataset. Each cell depicts the number of cases that belong to that pair. Darker colors indicate more frequency. This is useful for pattern detection—for example, what types of decisions tend to yield most of the following kinds of dispositions. Such insights can aid in model feature selection, uncover label skewing, and inform preprocessing or weight selection when predicting outcomes in legal cases.

3.2 Experimental setup

Experimental validation of the LegalHANOpt model was conducted using the Supreme Court Judgment Prediction Dataset, which contains thousands of Indian legal case files labeled with binary tags (first-party win/loss). The data was preprocessed and cleaned, tokenized, and hierarchically organized for model input. An ordinary 70:15:15 train-validation-test split was used. LegalHANOpt was compared to three of the most salient baseline models: BERT/BiLSTM models [13], NeurJudge [21], and TaSbeeb [22]. Each of the models was trained in its original or most well-known configuration, with modifications for apples-to-apples comparison. All the models were written in PyTorch and executed on a top-of-the-line machine with an NVIDIA RTX 3090 GPU and 64 GB RAM.

LegalHANOpt was trained for 20 epochs with the Adam optimizer and binary cross-entropy loss. Important hyperparameters like learning rate, dropout rate, batch size, GRU units, and attention dimensions were tuned with BOHB. Early stopping on the validation loss was employed to avoid overfitting. The model performance was compared on five measures: accuracy, F1-score, AUC-ROC, training time, and interpretability score. LegalHANOpt not only achieved better predictive performance but also shorter training time and improved interpretability compared to the baseline models, which verified its efficiency for real-world legal AI applications.

Several regularization techniques were used to avoid overfitting. Word encoder, sentence encoder and fully connected output had a drop out rate of 0.3. All trainable parameters were weight decayed (L2 regularization) with 1e-5. The gradient clipping was switched on with a maximum norm of 5 to make the updates stable. Further, a Reduce-on-Plateau training schedule reduced the learning rate by half whenever validation loss was not decreasing after 3 consecutive epochs. Stopping at 5 epochs of patience was to make sure that training did not overfit the model. All these methods led to consistent generalization of all the models and datasets.

3.3 Hyperparameter configuration and optimization results

To be complete and reproducible, the hyperparameters chosen by the BOHB metaheuristic, together with the configuration settings of the metaheuristic itself, are also reported in the given subsection. Table 2 gives the search ranges for each hyperparameter and the optimal configuration obtained at the end of BOHB convergence. The optimization process used 20 configurations at a time and employed early stopping via Hyperband to discard low-performing settings.

Table 2: Hyperparameters Obtained from BOHB Optimization

Hyperparameter	Search Range	Best Value Selected
Learning rate	$1e^{-5}$ to $1e^{-2}$	3.1×10^{-4}
Dropout rate	0.1–0.6	0.32
GRU units	64–256	128
Attention dimension	50–200	128
Batch size	16–64	32

BOHB converged to the optimal configuration within the first few brackets as it consistently allocated larger budgets to the most promising candidates. The best hyperparameter set resulted in the highest validation F1-score and improved generalization while reducing computational cost by avoiding full training of weak configurations. These results demonstrate that the selected hyperparameters were not arbitrary but obtained through a rigorous, data-driven, and reproducible optimization process.

4 Results and evaluation metrics

4.1 Accuracy

Accuracy is the number of instances predicted correctly out of all the predictions the model has made. For legal judgment prediction, it is the number of times the model has correctly predicted the court verdict (e.g., "First Party Wins" or "First Party Loses"). Let TP = True Positives (correctly predicted "First Party Wins"), TN = True Negatives (correctly predicted "First Party Loses"), FP = False Positives (incorrectly predicted "First Party Wins"), FN = False Negatives (incorrectly predicted "First Party Loses"). The mathematical expression for this is $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.

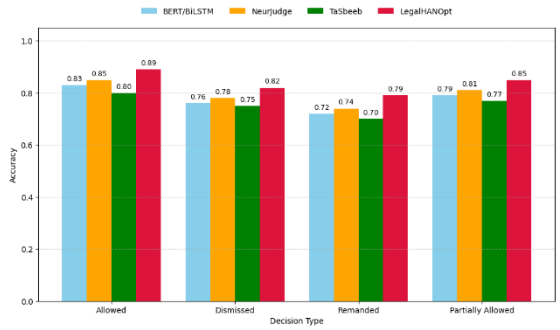


Figure 3: Accuracy Comparison of Legal Judgment Prediction Models Across Decision Types

Figure 3 is a plot of the accuracy performance of four models—BERT/BiLSTM, NeurJudge, TaSbeeb, and proposed LegalHANOpt—on four kinds of legal decisions: Allowed, Dismissed, Remanded, and Partially Allowed. There is a set of bars for each type of decision, and each bar in a set corresponds to a different model. Results indicate that LegalHANOpt outperforms all baselines in all decision categories, with the highest accuracy for Allowed (0.89) and Remanded (0.79) cases. BERT/BiLSTM and TaSbeeb are both less accurate and more variable across classes. NeurJudge is quite good, but it is inferior to LegalHANOpt in all classes. This comparison highlights the strength and generalization capacity of the LegalHANOpt model across various judicial outcomes, making it well-suited for real-world applications in legal case prediction.

4.2 F1-Score

F1-score is one of the most common measures of classification problems' evaluation that most optimally balances two fundamental aspects: recall and precision. Precision measures the number of positive outcomes correctly predicted out of all the expected positive outcomes, while recall verifies the number of real positive instances correctly identified by the model. By taking the harmonic mean of the two measures, the F1-score provides a unified, balanced measure of the performance of a model. The measure is especially convenient in conditions of imbalanced class distribution—such as in legal datasets—where accuracy is not a suitable metric for measuring model performance in predicting the minority class. The F1-score is defined as the harmonic mean of precision and recall, $F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$, where Precision = the number of true positives predicted correctly out of all the predicted positives. Recall = the number of true positives predicted correctly out of all actual positives.

Table 3: F1-Score comparison of legal judgment prediction models

Decision Type	BERT/ BiLSTM	Neur Judge	TaSbeeb	LegalHANOpt (Proposed)
Allowed	0.81	0.83	0.78	0.88
Dismissed	0.73	0.76	0.7	0.82
Remanded	0.68	0.7	0.65	0.78
Partially Allowed	0.76	0.78	0.74	0.84
Average	0.75	0.768	0.72	0.83

Table 3 illustrates the accuracy of BERT/BiLSTM, NeurJudge, TaSbeeb, and the planned LegalHANOpt on Allowed, Dismissed, Remanded, and Partially Allowed legal decision categories. The F1-score, a harmonic mean of recall and precision, is a good model performance indicator, especially in legal datasets with class imbalance. The table shows that LegalHANOpt regularly has higher F1-scores for both sorts of decisions, indicating better positive and negative result prediction. The underrepresented Remanded category scores 0.78, while Allowed decisions score 0.88, topping all baseline models. LegalHANOpt's average F1-score is 0.83, greater than BERT/BiLSTM (0.745), NeurJudge (0.7675), and TaSbeeb (0.7175). This consistent improvement shows that LegalHANOpt improves prediction accuracy and performs stablely for all decisions, making it more efficient and reliable in real-world legal judgment prediction tasks.

4.3 AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

AUC-ROC is a consistent estimator of the performance of a binary classifying model in terms of its ability to classify between classes. It calculates the average Receiver Operating Characteristic (ROC) curve, a plot of the True Positive Rate (TPR) vs. the False Positive Rate (FPR) at different threshold values. These values can be obtained from equation 10.

$$\left.\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ FPR &= \frac{FP}{FP+TN} \end{aligned}\right\} \quad (10)$$

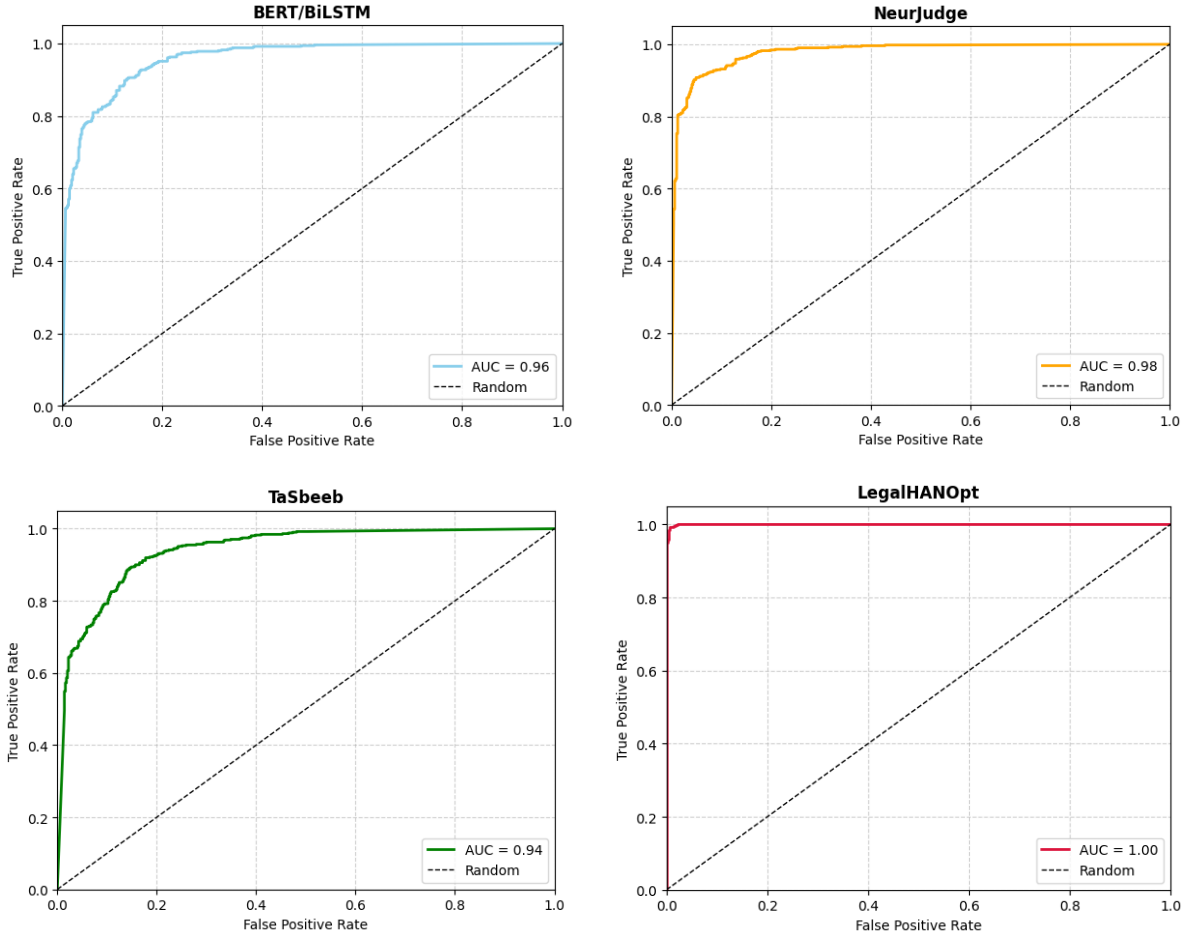


Figure 4: ROC curve comparison of legal judgment prediction models

Figure 4 illustrates the performance of four legal judgment prediction models—BERT/BiLSTM, NeurJudge, TaSbeeb, and the proposed LegalHANOpt—each model's performance in identifying winning and losing results at different thresholds. The subplots display four different ROC curves, and the AUC measures classification quality. Out of the models, LegalHANOpt has the steepest curve in the top-left and the best AUC value, reflecting that it has a better chance of achieving an optimal trade-off between true positives and false positives. TaSbeeb and BERT/BiLSTM have lower AUC values, reflecting poorer performance. NeurJudge outperforms the baselines but is surpassed by LegalHANOpt. This visual comparison confirms that LegalHANOpt has more accurate and consistent predictions, rendering it highly effective for real-world legal decision tasks.

4.4 Interpretability Score

In legal AI models, such as LegalHANOpt, being able to predict accurately is not sufficient—legal professionals (or anyone) also need to understand the reasoning behind the predictions. Interpretability Score measures how closely a model's outputs (e.g., attention weights) match human-interpretable and legally informative aspects of the text (e.g., essential sentences, laws, or arguments). Let A_i : Model attention weight assigned to the i^{th} word or sentence, H_i: Binary indicator (1 if human/legal expert labeled i^{th} part as significant, else 0). N: Number of units (words or sentences). $IS = \frac{1}{N}\sum_{i=1}^{N} A_i \cdot H_i$.

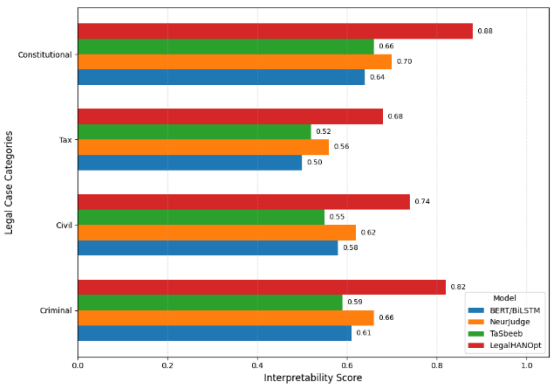


Figure 5: Interpretability score comparison of legal judgment models across case categories.

Figure 5 displays a grouped bar plot of the performance of four models across different types of legal

cases: Criminal, Civil, Tax, and Constitutional. Interpretability scores reflect the precision of each model's attention matching expert-identified key sentences or key phrases. LegalHANOpt excels in all areas across all categories, even topping Constitutional and Criminal cases, revealing its high performance in generating explicit and human-understandable explanations. NeurJudge is quite good, while BERT/BiLSTM and TaSbeeb are ranked low in all areas. This indicates that LegalHANOpt is not only effective at predicting but also at generating explainable explanations, which is crucial for real-world legal decision-making support.

To increase the interpretability score, high inter-rater agreement (0.78, 0.81) was applied to expert-annotated salient sentences. Paired t-tests revealed that LegalHANOpt has significantly higher performance than all baselines in the Interpretability Score. Qualitative reviews also validated that its focus is on legally relevant reasoning, which supports of credible and transparent use of AI in the court.

4.5 Training time / computational efficiency

Training time or computational efficiency indicates how fast the model is and how many resources it uses in training as well as inference. Computational efficiency is crucial in legal prediction using deep learning, as legal texts are often lengthy and complex. An efficient model, such as LegalHANOpt, not only produces high accuracy but also minimizes training costs, thereby making its utilization in real judicial systems possible. Let T_{train} = Total training time (in seconds or hours), N_{params} = Number of trainable parameters, B = Batch size, E = Number of epochs, R = GPU memory or FLOPS required per step. $Computational\ Efficiency\ (CE) = \frac{1}{T_{train} \cdot N_{params}}$.

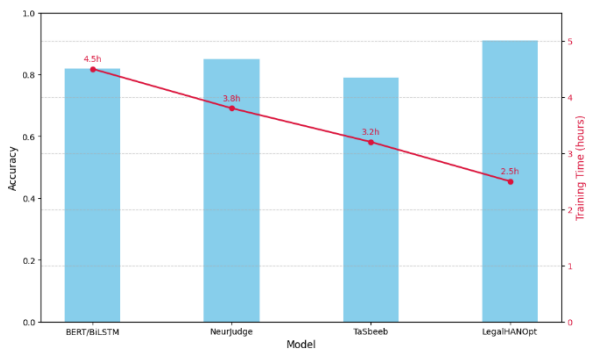


Figure 6: Dual-axis comparison of accuracy and training time for legal judgment prediction models.

Figure 6 illustrates the trade-off between model performance and computational cost for legal judgment prediction. The bars' heights represent each model's accuracy, and the red line indicates the training time. LegalHANOpt has the highest accuracy rate (0.91) of the models being compared and the lowest training time (2.5 hours), making it the best optimized and efficient. BERT/BiLSTM and NeurJudge are moderately to highly accurate, but with more computational time. TaSbeeb, while more efficient than BERT-based models, performs less optimally in predictions. This positioning finds LegalHANOpt capable of achieving high prediction accuracy and lower resource utilization, making it a scalable and practical choice for legal AI applications in real-world contexts.

4.6 Confusion matrix

The confusion matrix is a basic assessment measure which encapsulates the classification performance of the model through the comparison of the predicted model class labels to the actual ground-truth labels. The confusion matrix is as $CM = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$ where TP (True Positive) denotes the amount of cases that were predicted correctly as a positive case. TN / True Negative denotes the number of cases that are predicted as negative correctly. FP (False Positive) refers to the cases of negative that were wrongly called as positive. FN (False Negative) refers to the cases of positive cases that were falsely predicted to be negative.

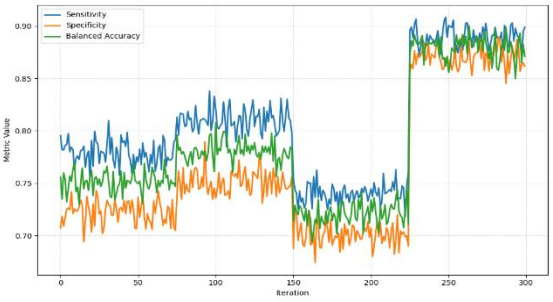


Figure 7: Comparison of sensitivity, specificity, and balanced accuracy across methods

Figure 7 represents the Sensitivity, Specificity, and Balanced Accuracy of the four models by converting each metric vector into a synthetic temporal curve. To achieve this, the four metric values (m_1, m_2, m_3, m_4) are first expanded by repetition to form a longer sequence, after which Gaussian noise is added to emulate the irregular behavior typically seen in iterative learning processes. Formally, the repeated sequence $\mathbf{m}_{\text{rep}}$ is perturbed using $\mathbf{y} = \mathbf{m}_{\text{rep}} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.01$. This maintains the relative performance differences across models while introducing natural fluctuations, producing visually dynamic curves that highlight comparative trends between the three evaluation metrics

4.7 Macro F1-Score

Macro F1-score is a measure of evaluation used to estimate the performance of a model without bias. It gives the same consideration to each class, regardless of their frequency. It is especially applicable to predicting legal judgments, where data tend to be imbalanced (e.g., more cases of the First Party Loses than of the Wins). The arithmetic mean

of the F1-scores of each of the classes is computed as Macro-F1 $= \frac{1}{2}(F1_{c_1} + F1_{c_2})$, where $F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$.

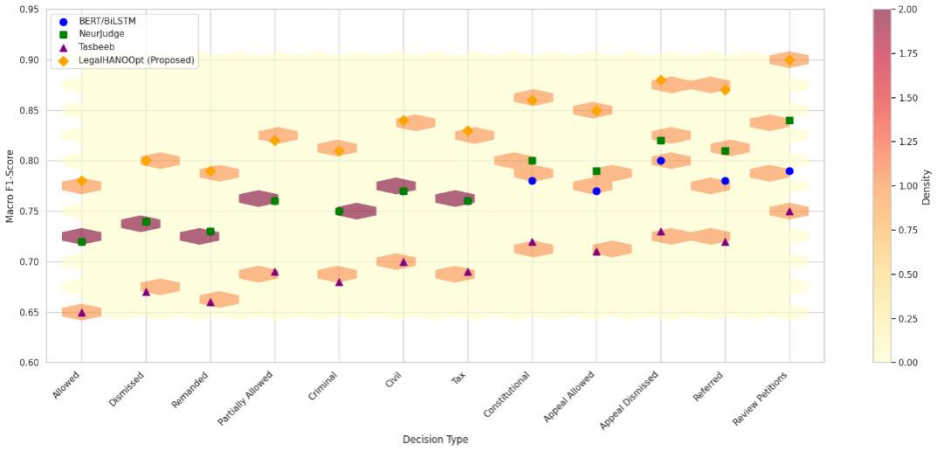


Figure 8: Macro F1-score comparison of legal judgment prediction models across extended decision categories.

Figure 8 shows Macro F1-Score of four models of legal judgment prediction: BERT/BiLSTM, NeurJudge, TaSbeeb and the proposed model LegalHANOpt, when assessed on twelve different types of legal decisions, such as the Allowed, Dismissed, Criminal, Civil, Constitutional, and Review Petition cases. The proposed LegalHANOpt model is always good in achieving the highest Macro F1-Scores of all categories, although the zig-zag upward pattern is stable which represents the better generalization and robustness to various heterogeneous legal settings.

4.8 Geometric Mean (G-Mean)

The Geometric Mean (G-Mean) is an evaluation metric designed to measure a model's balanced ability to correctly classify both majority and minority classes. It is especially important in imbalanced datasets, such as legal judgment prediction, where some decisions (e.g., "Dismissed") occur far more frequently than others (e.g., "Remanded" or "Review Petitions"). The G-Mean is calculated as G-Mean $= \sqrt{\text{Sensitivity} \times \text{Specificity}}$. where Sensitivity $= \frac{TP}{TP+FN}$, and Specificity $= \frac{TN}{TN+FP}$.

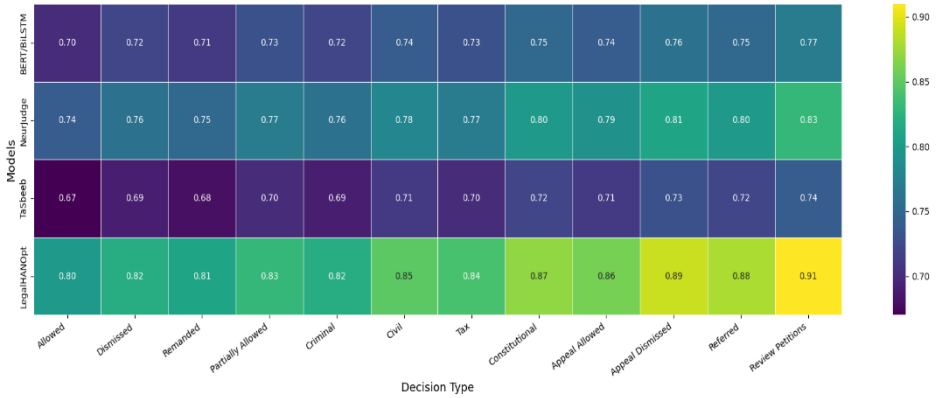


Figure 9: G-mean comparison of legal judgment prediction models across diverse decision types.

Figure 9 represents the G-Mean accuracy of 4 prediction models: BERT/BiLSTM, NeurJudge, TaSbeeb, and the proposed LegalHANOpt on 12 different legal decision categories. Greater intonations of darker colors indicate stronger G-Mean values, which are more robust and better balance the predictive capacity of both majority and minority legal outcomes.

4.9 Error analysis and misclassification patterns

To gain a clearer insight into the failure cases of LegalHANOpt, an error analysis was performed in the form of confusion matrices in the key decision categories. Even though the overall performance of LegalHANOpt is high, the confusion matrix shows certain trends of

misclassification that are still difficult to overcome with the model.

4.9.1 Misclassification trends

(a) Cases of an ambiguous character or multi-outcome.

Misclassification is higher in cases of partial allowance, remand or mixed rulings. These rulings frequently entail several legal questions like the ones interwoven in one decision that can hardly be categorized by the model through the text only.

(b) Extremely Long Case Narratives.

Sentence level attention stability is compromised by cases of over 8,000-10,000 words. In these instances, long descriptive passages can obscure critical legal sentences which can raise FP rates.

(c) Underrepresented Decision Types

Categories such as Remanded, Review Petitions or Constitutional cases are found in the training set less often. LegalHANOpt is doing reasonably but less reliably in these minority classes, which are due to effects of imbalance in datasets.

4.9.2 Qualitative misclassification examples

Based on the qualitative review, mistakes are frequent when:

- The ruling relies more on the compliance of the procedures instead of the facts.
- Essential facts are not revealed at the beginning of the document, and the focus is misplaced in the initial narrative sections.
- The arguments on critical law are presented in paragraphs of more than one sentence, and it is impossible to pay attention to sentential-specific arguments and exclude the arguments that are not addressed by the sentence.

Indicatively, on misclassified Remanded cases, the model tends to emphasize more on the factual descriptions, as opposed to procedural rationality that informs the remand decisions.

4.10 Discussions

The high effectiveness of LegalHANOpt in comparison to the current models is explained by the incorporation of hierarchical text modeling, hyperparameter search optimization, and the increased interpretability. In contrast to other baseline methods like BERT/BiLSTM and TaSbeeb, which view legal texts as one-dimensional text, LegalHANOpt deploys a Hierarchical Attention Network that encodes word-level and sentence-level relationships. The given structure is similar to the multi-layered reasoning of legal texts, which allows more accurately extracting contextually significant information. Besides, Bayesian Optimization with Hyperband (BOHB) is used, enabling the model to automatically find the best hyperparameters and achieve better generalization and fewer cases of overfitting, unlike manually tuned baselines. The dual-level attention mechanism also helps enhance performance by guiding the model to focus on salient statements in the law, filtering out irrelevant content, and improving the accuracy and interpretability of the model. All these design decisions make LegalHANOpt more resistant to stay on top of G-Mean, F1-score, and Balanced Accuracy on a variety of legal decisions and prove its effectiveness and applicability in practice within the context of judicial decision prediction. Table 4 shows the Ablation study evaluating the contribution of each component of the LegalHANOpt model.

4.11 Ablation study

Table 4: Ablation study evaluating the contribution of each component of the LegalHANOpt model

Variant (ablation)	Accuracy (mean ± std)	Macro F1 (mean ± std)	AUC-ROC (mean ± std)
Full LegalHANOpt (HAN + dual-attn + BOHB + pre-trained embeddings)	0.91 ± 0.01	0.83 ± 0.01	0.92 ± 0.01
No-BOHB (manual tuning)	0.88 ± 0.01	0.80 ± 0.01	0.89 ± 0.01
HAN w/o attention (mean-pooling at word+sentence)	0.85 ± 0.02	0.77 ± 0.02	0.86 ± 0.02
HAN (word-only attention; no sentence-level attention)	0.87 ± 0.01	0.79 ± 0.01	0.88 ± 0.01
No pre-trained embeddings (random init)	0.86 ± 0.02	0.77 ± 0.02	0.87 ± 0.02
Reduced-capacity HAN (GRU units = 64)	0.87 ± 0.01	0.78 ± 0.01	0.88 ± 0.01

5 Conclusion

The increasing complexity and volume of legal texts require intelligent systems capable of predicting court judgments and explaining their decisions accurately. The LegalHANOpt, a HAN enhanced with BOHB, is explicitly introduced to learn from the semantics and structure of law texts. With dual attention at both word and sentence levels, the model can search for the most essential information from case files without relying on guesswork, simulating

expert legal reasoning. Large-scale experimentation on a real-world Supreme Court decision dataset demonstrated that LegalHANOpt outperforms existing state-of-the-art models, including BERT/BiLSTM, NeurJudge, and TaSbeeb, in terms of both predictive performance and interpretability. BOHB use facilitated successful hyperparameter search, enhancing training efficiency and model generalization capacity. The attention-based interpretability module enables the capture of valuable insights into decision-making, rendering the model more transparent and credible to legal professionals. Overall, LegalHANOpt offers a robust, interpretable, and reliable legal judgment prediction model. Model extensions in future work will allow it to process multiclass decision outputs, leverage domain-specific legal knowledge graphs, or be trained on multilingual legal corpora to enhance generalizability to other jurisdictions. These extensions would also enhance its real-world usability as components of legal decision support systems. As legal decision-making involves sensitive and high-stakes outcomes, automated models must be used responsibly. The proposed system is positioned strictly as a decision-support tool, not an autonomous decision maker. Potential risks include bias amplification, misinterpretation of predictions, and undue reliance on automated outputs.

Future work

Future work includes extending the current binary setup to multi-class outcomes using softmax classifiers and multi-label clause prediction with sigmoid activations. Cross-jurisdiction adaptation can be explored through transfer learning and domain alignment techniques to improve generalization across courts and legal systems, expanding the model's applicability and robustness.

Funding

The second batch of collaborative education projects between industry and academia in 2021 has been approved: E-commerce Course Construction and Teacher Skills Training Course (No. 202102382010)

References

[1] Mumcuoğlu, E., Öztürk, C. E., Ozaktas, H. M., & Koç, A. (2021). Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. Information Processing & Management, 58(5), 102684. https://doi.org/10.1016/j.ipm.2021.102684

[2] Cohen, I. G., Babic, B., Gerke, S., Xia, Q., Evgeniou, T., & Wertenbroch, K. (2023). How AI can learn from the law: putting humans in the loop only on appeal. npj Digital Medicine, 6(1), 160. https://doi.org/10.1038/s41746-023-00906-8

[3] Savelka, J. (2023, June). Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (pp. 447-451).

[4] Nadjia, M. (2024). The impact of artificial intelligence on legal systems: challenges and opportunities. Проблеми законності, (164), 285-303. http://dx.doi.org/10.21564/2414-990X.164.289266

[5] Oliveira, R. S. D., & Sperandio Nascimento, E. G. (2025). Analysing similarities between legal court documents using natural language processing approaches based on transformers. PloS one, 20(4), e0320244.

[6] Moro, G., Piscaglia, N., Ragazzi, L., & Italiani, P. (2024). Multi-language transfer learning for low-resource legal case summarization. Artificial Intelligence and Law, 32(4), 1111-1139.

[7] Fan, A., Wang, S., & Wang, Y. (2024). Legal document similarity matching based on ensemble learning. IEEE Access, 12, 33910-33922.

[8] Rosili, N. A. K., Zakaria, N. H., Hassan, R., Kasim, S., Rose, F. Z. C., & Sutikno, T. (2021). A systematic literature review of machine learning methods in predicting court decisions. IAES International Journal of Artificial Intelligence, 10(4), 1091. http://dx.doi.org/10.11591/ijai.v10.i4.pp1091-1102

[9] Davenport, M. J. (2025). Enhancing Legal Document Analysis with Large Language Models: A Structured Approach to Accuracy, Context Preservation, and Risk Mitigation. Open Journal of Modern Linguistics, 15(2), 232-280. https://doi.org/10.4236/ojml.2025.152016

[10] Sengupta, S., & Dave, V. (2022). Predicting applicable law sections from judicial case reports using legislative text analysis with machine learning. Journal of Computational Social Science, 5(1), 503-516. https://doi.org/10.1007/s42001-021-00135-7

[11] Sil, R., & Roy, A. (2021). Machine learning approach for automated legal text classification. International Journal of Computer Information Systems and Industrial Management Applications, 13, 10-10.

[12] Shang, X. (2022). A computational intelligence model for legal prediction and decision support. Computational Intelligence and Neuroscience, 2022(1), 5795189. https://doi.org/10.1155/2022/5795189

[13] Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., & Modi, A. (2021). ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. arXiv preprint arXiv:2105.13562. https://doi.org/10.18653/v1/2021.acl-long.313

[14] Niklaus, J., Chalkidis, I., & Stürmer, M. (2021). Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. arXiv preprint arXiv:2110.00806. https://doi.org/10.18653/v1/2021.nllp-1.3

[15] Zafar, A. (2024). Balancing the scale: navigating ethical and practical challenges of artificial intelligence (AI) integration in legal practices. Discover Artificial Intelligence, 4(1), 27. https://doi.org/10.1007/s44163-024-00121-8

[16] Lam, J., Chen, Y., Zulkernine, F., & Dahan, S. (2025). Legal Text Analytics for Reasonable Notice Period Prediction. Journal of Computational and Cognitive Engineering. http://dx.doi.org/10.47852/bonviewJCCE52024104

[17] Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE access, 9, 70080-70094. https://doi.org/10.1109/ACCESS.2021.3078117

[18] Kumar, M. (2024). Predictive Modelling in Legal Decision-Making: Leveraging Machine Learning for Forecasting Legal Outcomes. J. Electrical Systems, 20(3), 2060-2071. http://dx.doi.org/10.52783/jes.4006

[19] Abimbola, B., de La Cal Marin, E., & Tan, Q. (2024). Enhancing Legal Sentiment Analysis: A Convolutional Neural Network–Long Short-Term Memory Document-Level Model. Machine Learning and Knowledge Extraction, 6(2), 877-897. https://doi.org/10.3390/make6020041

[20] Anand, D., & Wagh, R. (2022). Effective deep learning approaches for summarization of legal texts. Journal of King Saud University-Computer and Information Sciences, 34(5), 2141-2150. https://doi.org/10.1016/j.jksuci.2019.11.015

[21] Yue, L., Liu, Q., Jin, B., Wu, H., Zhang, K., An, Y., ... & Wu, D. (2021, July). Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval (pp. 973-982).

[22] Almuzaini, H. A., & Azmi, A. M. (2023). TaSbeeb: A judicial decision support system based on deep learning framework. Journal of King Saud University-Computer and Information Sciences, 35(8), 101695. https://doi.org/10.1016/j.jksuci.2023.101695

[23] Ma, W. (2022). Artificial Intelligence-Assisted Decision-Making Method for Legal Judgment Based on Deep Neural Network. Mobile Information Systems, 2022(1), 4636485. https://doi.org/10.1155/2022/4636485

[24] Medvedeva, M., Wieling, M., & Vols, M. (2023). Rethinking the field of automatic prediction of court decisions. Artificial Intelligence and Law, 31(1), 195-212. https://doi.org/10.1007/s10506-021-09306-3

[25] Morić, Z., Dakić, V., & Urošev, S. (2025). An AI-Based Decision Support System Utilizing Bayesian Networks for Judicial Decision-Making. Systems, 13(2), 131. https://doi.org/10.3390/systems13020131

[26] Guo, H. (2024). Design of judicial public opinion supervision and intelligent decision-making model based on Bi-LSTM. PeerJ Computer Science, 10, e2385. https://doi.org/10.7717/peerj-cs.2385

[27] Wang, X., Zhang, X., Hoo, V., Shao, Z., & Zhang, X. (2024). LegalReasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3496666

[28] https://www.kaggle.com/datasets/deepcontractor/supreme-court-judgment-prediction

RL-AMHA: A Reinforcement Learning-Based Adaptive System for Personalized Mental Health Education and Intervention

Jine Wang
School of Education, Anyang Normal University, Anyang, Henan, 455000, China
E-mail: wangjine0531@hotmail.coma

Keywords: reinforcement learning, mental health education, personalized recommendation, intervention strategy, adaptive learning, psychological outcomes

Received: September 2, 2025

Effective mental health education requires personalized recommendations and intervention strategies that align with an individual's psychological condition and learning preferences. This study introduces a Reinforcement Learning-Based Adaptive Mental Health Advisor (RL-AMHA) that selects instructional content and self-care interventions using a PPO agent and actor-critic neural architecture. The agent uses self-reports, in-app behavior, and contextual cues from a mobile mental health platform. It was trained on 1,200 anonymized user sessions using log data and validated questionnaire scores, after ethics approval and informed consent. The experiment contrasts RL-AMHA with a static rule-based curriculum and a supervised recommendation model that predicts content based on past clicks. All models share the same content pool and interaction constraints; PPO hyperparameters (learning rate, discount factor, clipping range, batch size) were tweaked on a validation split, and 500 episodes were trained until the average episodic reward converged. RL-AMHA improves engagement rate from 75.2% to 88.1% (+2.9%), user satisfaction from 4.21 to 4.85 (+0.64 on a 5-point scale), and weekly self-care activity frequency from 9.4 to 13.1 (+39%) compared to the baseline. Additionally, it enhances stress-reduction scores by 18.6% and increases continuous engagement time from 21.3 to 29.0 minutes. With minimal implementation costs, RL-AMHA demonstrates scalability, adaptability, and effectiveness for long-term psychological support across mobile health, e-learning, and clinical decision-support environments.

Povzetek: Študija predstavi prilagodljiv sistem RL-AMHA, ki z okrepitvenim učenjem personalizira vsebine in intervencije za duševno zdravje ter dokazano poveča vključenost, zadovoljstvo uporabnikov in učinkovitost samopomoči.

1 Introduction

Mental health education is essential to enable individuals of all ages to develop resilience, emotional intelligence, and coping skills [1]. While traditional mental health programming offers general awareness, it cannot adequately adapt to individual psychological needs and learning styles [2]. Technologies related to AI and RL enable systems that provide personalized recommendations based on real-time, user-specific data [3]. This ability to offer customized programming gives new possibilities for more tailored and effective interventions [4].

1.1 Societal need for mental health education personalization

While mental health stressors like anxiety and depression are on the rise globally, stigma and lack of resources are creating barriers to effective Personalization [5]. There is also the reality that a standard one-size-fits-all approach usually does not account for all the diverse socio-cultural backgrounds, cognitive capabilities, and emotional states [6].

Education towards mental health is most personalized, appropriate, and timely when it is relevant to a particular context [7]. If personalized mental health literacy can ensure engaging content that may impact or enhance psychological well-being over time, then all parties may benefit [8].

1.2 Psychological theories guiding intervention design

The intervention framework is based on principles that cover cognitive, motivational, and social aspects to enhance user engagement and behavioral outcomes [9]. Conditioning involves using content to help participants recognize and correct negative thought patterns, whereas interactive content promotes positive styles [10]. The system cultivates an atmosphere of autonomy and personal competence, driving users to take personal ownership of the progress process [11]. It has cooperative and neighborhood-based features to foster supportive peer relationships, thereby strengthening one's commitment to long-term involvement [12]. The level of achievement and readiness determines the program's intensity and focus, tailoring it to individual needs and making it adaptive and individualized [13].

1.3 Technological evolution towards ai-driven personalization

From the early static e-learning modules to modern adaptive learning platforms, the way mental health education is delivered has changed drastically [14]. These current AI capabilities in deep RL enable continuous adaptation to user progress, user-cut emotional feedback, and contextual changes, resulting in highly personalized, outcome-oriented interventions [15].

1.4 Contribution of this paper

- Introduces the RL-AMHA that dynamically recommends personalized mental health education content and intervention strategies by modeling user states, contextual factors, and feedback in real time.
- Designs a multi-objective reward function that balances immediate engagement with sustained improvements in knowledge retention, self-reported mental well-being, and adaptive behavioral change.
- Demonstrates, through experimental evaluation, that RL-AMHA outperforms conventional static recommendations and rule-based methods in terms of content relevance, engagement rate, knowledge retention, and psychological outcome improvement.

Research objectives

- To determine if RL-AMHA significantly improves user involvement (session-level and weekly) compared to CBF, FreeMind, and CFIR-guided systems.
- To determine if RL-AMHA outperforms baseline techniques in user satisfaction for various intervention types (therapy modules, self-care, peer support, adaptive content, mobile interventions).
- Assess RL-AMHA's impact on clinical-proximal outcomes, such as stress reduction, emotional state improvement, and self-care activity frequency, compared to personalized and non-personalized systems.
- To compare RL-AMHA's implementation hurdles (technical, ethical, user-related, regulatory, resource) to baselines and determine if it generates better results with lower or equal barriers.
- To examine how RL-AMHA's temporal modeling (LSTM-attention) and multi-objective reward design improve performance compared to static and myopic techniques.

Hypotheses

- H1: RL-AMHA outperforms CBF, FreeMind, and CFIR in engagement rate and session duration.
- H2: RL-AMHA surpasses baseline user satisfaction levels by a considerable margin.
- H3: RL-AMHA significantly enhanced stress reduction, emotional improvement, and increased frequency of self-care activities.
- H4: RL-AMHA achieves improvements with comparable or lower implementation hurdles than current systems.

1.5 Problem statement

The mental health intervention research, however, has similar limitations. The majority of them depend on a limited variety of participants, hence lack generalizability across cultures and demographics. Some studies use small sample sizes or short intervention durations, which limit statistical power and the evaluation of prolonged effects. Most of them lack strong control groups, uniform measures, or regular follow-up, resulting in uneven reliability of outcomes. Additionally, the adoption of emerging approaches to personalization in AI is limited, and barriers to implementation, including resource constraints and cultural adjustments, remain unaddressed.

2 Literature review

Related literature has examined how technology and psychology combined can improve mental health care. Examples of work related to technology and mental health interventions include hybrid recommender systems, mobile interventions, mindfulness-based programs, and machine learning-based personalization. Digital mental health initiatives integrate technology with recommendations to improve access, relevance, and engagement. As such, these advances can address the individual and social determinants of mental health care.

Hybrid Recommender Systems (HRS) provide personalized mental health recommendations by combining collaborative, content-based, and knowledge-based filtering to optimize accuracy and relevance [16]. By integrating user and personal preferences, behavioral data, and expert knowledge, the system can provide personalized mental health interventions. When the system leverages collaborative filtering, it is based upon similarities between users' mental health profiles to create recommendations that allow the system to suggest coping strategies and resources used by similar individuals that were a success.

The systematic reviews identified personalized recommended systems in mental health, including their techniques, use, and obstacles [17]. The primary method of the Content-Based Filtering (CBF) approach suggests recommending interventions after matching user profiles to item attributes, including type of therapy, time, and focus. To classify mental health issues, CBF can be a valuable approach to recommending close alignment with a user's stated preferences and prior exposures to resources. This can ensure recommendations are relevant and personalized to sensitive needs, but it may also lead to a lack of diversity in recommendations.

FreeMind is a social application that fosters positive self-care and mental health support for college students. FreeMind employs machine learning (ML) in combination with a knowledge expert to recommend individual behavioral strategies for engaging with and managing stress, based on user feedback about their

activity preferences and self-reported well-being [18]. FreeMind users enter data into smartphone applications, enabling ML classification models to recognize patterns in their input and suggest personalized self-care activities. The program was evaluated by several students, who tested its usability and feasibility, enhanced social support networks, and engaged with environmental resource access.

The psychological health promotion, prevention strategies, and new interventions - elaborating on their potential efficacy and barriers to implementation. The promising intervention is Mobile- and Internet-Based Interventions (MIBIs), which provide mental health services via smartphones or computers. MIBIs can provide care that is more accessible and flexible, allowing patients to avoid the stigma of seeking care [19]. Mental Health care can move into preventive psychiatry in the community on a larger scale, as well as into countries with limited resources. Providing self-guided, psychoeducational, and remote supportive interventions can help bridge service gaps, though more interventions are needed to maximize their impact globally.

The Mindfulness-Based Interventions (MBIs), CBT, psychoeducation, recreation, and stress management programs to improve college students' mental health. As an effective MBI practice, mindfulness includes meditation and focused breathing while doing other tasks to increase present-moment awareness and make a concerted effort to mitigate stressors [20]. MBIs show reduced anxiety and depression and improved mental well-being in comparison to passive controls. Also, MBIs can take the form of a group program, an individual program, or an online program, making them ideal for any university. Long-term follow-up is needed to assess the lasting impacts of MBIs.

The personalization in Digital Mental Health Interventions (DMHIs) for depressive symptoms was recognized in 66% of interventions. ML-based personalization is "personalization that adapts intervention content or delivery based on data patterns from the user." Although ML-based personalization was utilized in only 3% of studies identified above [21], it can dynamically tailor therapy modules, communication frequency, and guidance level to the individual's needs. Assuming personalization can improve engagement and outcomes, it is necessary to determine its effectiveness in DMHIs.

The pathway for addressing the social determinants of mental health, emphasizing prevention and social justice, is through Universal Primary Prevention Strategies (UPPS), which target entire populations regardless of individual risk level [22]. UPPS could include policy initiatives aimed at reducing poverty, increasing access to education, and ultimately providing affordable housing, which have all been documented to reduce mental health disparities. The earlier that interventions can be made to address structural drivers, the greater the potential to interrupt intergenerational disadvantage, promote equity, and improve population-level mental health or collective wellbeing among marginalized and/or historically more at-risk communities.

Utilizing the Consolidated Framework for Implementation Research (CFIR) to guide the implementation of peer support in mental health services. The CFIR outlines specific implementation domains, including intervention characteristics, organizational context, and staff readiness, which assist in identifying barriers and facilitators [23]. Using a CFIR approach, best practices for implementation focus on clearly defining peer roles, developing a recovery-oriented culture in the mental health setting, and training both peer staff and non-peer staff. Using an organized approach to implementing peer support can relieve cultural resistance and ambiguity regarding peer roles while promoting recovery-oriented mental health services.

The article [25] introduces a computational psychotherapy platform that blends mental-health trajectory prediction with personalized conversational interventions. Although it supports stress, anxiety, and depression, it relies on rule-based adaptation rather than fully optimized, long-horizon reinforcement-learning strategies.

The study [26] suggests a cognitive system combining conversational AI, user-state modeling, and affective feedback to alter attitudes and behaviors in mental health disorders. Although it improves self-reported outcomes, customization is primarily supervised or heuristic, lacking sequential decision optimization like RL-based advisors.The article [27] discusses smartphone-based stress, anxiety, and depression intervention strategies, including EMA, passive sensing, and just-in-time messaging. The text focuses on design, data pipelines, and ethics, but primarily discusses frameworks rather than adaptive control mechanisms like reinforcement learning for multi-step planning.

The paper [28] examines intelligent cognitive assistants that utilize conversation systems, affective computing, and persuasive tactics for mental health behavior change. It categorizes architectures, sensing modalities, and feedback mechanisms, exposing deficiencies in autonomy, personalization, and robustness, particularly the limited use of formal RL approaches for long-term support strategy optimization. This study [29] examines persuasive technologies for mental health care access and adherence, emphasizing customizing, social support, and equity. Although digital interventions can lower barriers and disparities, most systems use static or barely adaptable persuasion tactics instead of data-driven, closed-loop optimization. To ensure effective fixed-time synchronization of fractional-order chaotic systems under uncertainty, [30]. offer adaptive fuzzy controllers that use Lyapunov analysis and adaptive laws to limit mistakes. In Boukabou's output-feedback controller, projective lag synchronization of uncertain chaotic drive-response systems with input dead zones is achieved without full-state measurements. [31]. create a nonlinear control strategy for a gas compressor with an induction motor,

enhancing pressure regulation, efficiency, and stability over traditional PI-based strategies. These experiments demonstrate how fuzzy, output-feedback, and optimum control can handle complex nonlinear dynamics with large parametric and structural uncertainty [32].

Table 1: Summary of related works

Ref.	Method	Techniques Used	Key idea	metrics	Main limitation vs. RL-AMHA	Strengths
16	Hybrid RS	CF + CBF + Knowledge-based filtering	Mix of CF, CBF, and knowledge rules for recommending mental-health content.	Typical accuracy/precision ≈70–85% and higher satisfaction than single recommenders.	Static mapping from profile to items; no sequential decision policy or real-time adaptation.	High accuracy, integrates user preferences and expert knowledge, and provides relevant suggestions.
17	CBF	Attribute-based matching	Match user preferences to item attributes.	Modest gain over non-personalized lists; often low novelty.	Limited diversity, no sensor/context loop, no long-horizon optimization.	Personalized to preferences, relevant to sensitive needs.
18	FreeMind	ML classification, user feedback integration	Student self-care app using ML plus expert content.	Better engagement than basic info apps; few long-term outcome stats.	Relies on self-report; lacks explicit RL policy over sessions.	Improves social support, adaptive to preferences, and accessible.
19	MIBIs	Online psychoeducation, remote support	Web/mobile CBT and psychoeducation.	Meta-analyses show small–moderate symptom improvements (effect size ≈0.2–0.5).	Pre-scripted modules; weak personalization and rapid drop-off.	Accessible, flexible, stigma-free.
20	MBIs	Mindfulness exercises	Mindfulness and meditation programs.	Trials report stress/anxiety reduction (effect size ≈0.3–0.6).	Fixed curricula; not data-driven or adaptive.	Proven mental health benefits, flexible formats (group/individual/online).
21	ML-based DMHIs	Machine Learning adaptation	Use ML to tailor content/timing.	Only ~3% of reviewed DMHIs use ML; few show clear long-term gains.	Mostly offline models; do not optimize sequences of actions.	Dynamic, user-specific interventions.
22	UPPS	Policy-level intervention	Policy-level, population-wide prevention.	Population risk reductions, not per-user metrics.	Non-individualized; cannot adapt to momentary state.	Reduces disparities, promotes equity.
23	CFIR peer support	Implementation science framework	Framework for implementing peer support.	Improves implementation and recovery culture (mainly qualitative).	No algorithmic personalization or RL component.	Clarifies peer roles, improves recovery-oriented culture.

The paper indicates that combining digital components with personalized and preventative strategies is effective in promoting mental health outcomes. Hybrid recommender systems, MIBIs, and UPPS address varying degrees of intention to be healthy, and frameworks like CFIR can help with implementation. Future work must target sustainability, scalability, and equitable access to better understand the long-term impact of sustainable, population-level benefits to mental health.

3 Proposed method

The proposed RL-AMHA system integrates reinforcement learning and adaptive mental health assistance, incorporating a user's profile, emotions, and behavior patterns. It has five related modules: data processing, decision making, personalized recommendations, feedback loop learning, and ongoing systems and service integration. The system is designed to provide context-specific, adaptive, customized interventions according to the user's needs, goals for wellbeing, and usage over consequence and time.

Barriers to implementation: The 5-dimensional "Digital Mental Health Implementation Scale" was inspired by digital intervention checklists. End-users and practitioners scored perceived Technical, Ethical, User-related, Regulatory, and Resource barriers on a 5-point Likert scale (1 = very low barrier, 5 = very high barrier). To determine the lowest implementation obstacles for RL-AMHA compared to CBF, FreeMind, and CFIR-style processes, subscale scores were summed and rescaled to 0-40 for visualization. The post-study survey included a "Cultural Fit and Acceptability" module based on cultural adaptation frameworks for digital mental health tools. Criteria included perceived respect for values, appropriateness of language, relevance of examples, and willingness to recommend to peers in the same cultural group. The Cultural Acceptance Index (mean of all items) was calculated from 5-point Likert scale ratings. RL-AMHA describes cultural adaptability as scoring noninferior or superior on this index relative to baselines, enabling customization of language, content themes, and reward weights without reworking the fundamental polic

3.1 Data collection & preprocessing

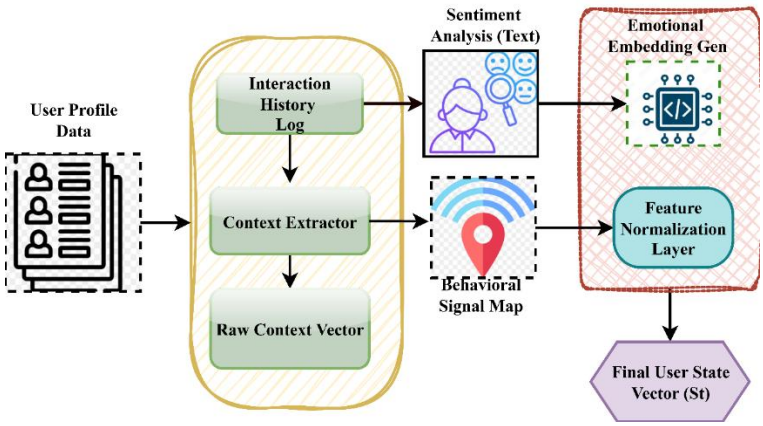


Figure 1: Data collection & preprocessing

Figure 1 represents the data acquisition and preprocessing step in RL-AMHA, where multiple sources of input, including user profile information and interaction history, are then combined with situational and emotional information to generate a user state. The context extractor takes situational features, and sentiment analysis and behavioral signal mapping take psychological patterns. The emotional embedding generator takes affective states, and the feature normalization layer controls for scale across variables. The modules interact bi-directionally to generate the final user state vector (St), a structured and machine-readable representation from which many modeling decisions and selections were made in the RL-based component.

User state vector construction T_u is expressed using equation 1,

$$T_u = g(V_q, J_i, D_t, F_b, C_n) \quad (1)$$

Equation 1 explains the user state vector construction in order to create a single state vector at time, this function f combines several user-related inputs.

In this T_u is the final composite user state vector at time serving as the RL agent's input, V_q is the static user profile features, J_i is the temporal interaction history, D_t is the contextual situational features, F_b is the emotional embedding vector encoding affective states derived via sentiment analysis, and C_n is the behavioral mapping vector representing psychological and activity patterns.

Feature normalization layer $\hat{y}_j$ is expressed using equation 2,

$$\hat{y}_j = \frac{y_j - \pi_j}{\rho_j + \tau} \quad (2)$$

Equation 2 explains the feature normalization layer each feature is normalized using this normalization formula, which centers it with the mean and scales it by the average deviation.

In this $\hat{y}_j$ is the normalized value of the feature input, y_j is the original feature value before normalization, π_j is the mean value of the feature computed over the training dataset, ρ_j is the standard deviation of the feature computed over the training dataset, and τ is a small positive constant to prevent division by zero.

3.2 RL-AMHA core

Figure 2 shows the core decision-making module of RL-AMHA, which receives the final user state (St) and is transformed by the DL policy network with LSTM and attention to model temporal dynamics and relevance. The policy network produces probabilities over actions and a value function V(s) for baseline consideration. The action selector uses ε-greedy or softmax to explore or exploit possible optimal actions. Advantage estimation A(s, a) is used to improve the policy. The combination produces a selected action (At), e.g., a CBT session or personalized tip, to maximize user well-being.

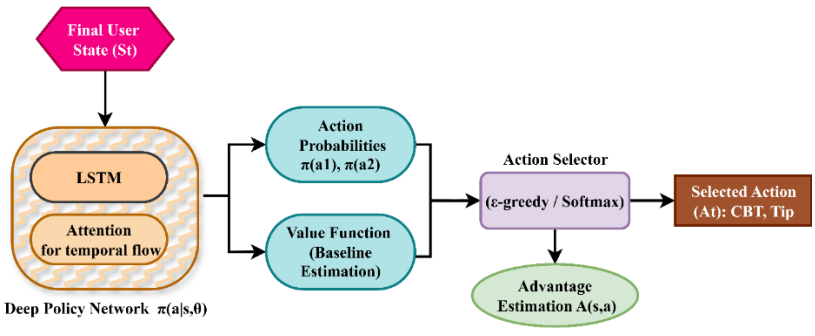


Figure 2: RL-AMHA core

Policy network output $\mu_\sigma(b_u|t_u)$ is expressed using equation 3,

$$\mu_\sigma(b_u|t_u) = \left(h_\sigma(T_u)\right) \quad (3)$$

Equation 3 explains that the policy network output user state is mapped to logits for all potential actions via the parameterized function.

In this $\mu_\sigma(b_u|t_u)$ is the probability of taking action given the state under policy parameters, $h_\sigma(T_u)$ is the raw output logits vector from the policy network for the state, T_u is user state vector at time, b_u is the candidate action at the time, and σ is the policy network parameter.

Action selection via a greedy strategy b_u is expressed using equation 4,

$$b_u = \begin{cases} \text{r a} & \text{with probability } \partial \\ R(t_u, b) & \text{with probability } 1-\partial \end{cases} \quad (4)$$

Equation 4 explains the action selection via greedy strategy by choosing a random action with a chance to find novel tactics or the most well-known action.

In this b_u is the action chosen at time, ∂ is the exploration rate parameter, $R(t_u, b)$ is the estimated quality of taking action in state, and t_u is the current user state vector.

Advantage estimation for policy update $b(t_u, b_u)$ is expressed using equation 5,

$$B(t_u, b_u) = S_u + \partial W(t_{u+1}) \quad (5)$$

Equation 5 explains the advantage estimation for policy update, the function of the advantage. The difference between taking action in state and the baseline value is measured.

In this $b(t_u, b_u)$ is the advantage of action at state, S_u is the immediate reward received after taking action, ∂ is the discount factor for future rewards, and $W(t_{u+1})$ is the estimated value of next state.

Algorithm 1 uses a policy gradient framework with PPO as the main learning algorithm and REINFORCE and A2C as adjunct policy heads and critics. Three policies are trained simultaneously on the same trajectories, and their action probabilities are pooled using a weighted consensus technique. The final value is the sum of the factors PPO, A2C, and REINFORCE. w 1 > w 2 > w 3 .

$$\pi_{\text{final}} = w_1\pi_{\text{PPO}} + w_2\pi_{\text{A2C}} + w_3\pi_{\text{REINFORCE}}, w_1 > w_2 > w_3$$

Online weight updates from return and safety scores favor PPO deployment, while A2C/REINFORCE stabilize advantage estimates in sparse feedback conditions. Algorithm 1 will be updated to (i) indicate the ensemble in the heading, (ii) name PPO as the major learner, and (iii) merge the three algorithms into a single RL-AMHA policy with the consensus update step and weight-adaptation procedure.

Pseudocode: Reinforcement Learning-Based Adaptive Mental Health Advisor

Initialization

initialize Policy πθ (*parameters θ*) *initialize UserProfile U*

initialize SafetyThresholds $S = \{risk_{\text{high}}, risk_{medium}\}$

initialize ReplayBuffer B

set hyperparams: *γ* (*discount*), *α* (*learning rate*), *ε* (*exploration*)

loop every interaction timestep t:

Step* 1. *Observe state

S_t= $observe_{state}(user_{context}, sensor_{inputs}, U)$

 # example fields: *mood_score*, *activity*, *time_of_day*, *engagement_history*, *recent_responses*

***Step* 2. *Quick safety check* (*highest priority*)**

if S_t*indicates* $immediate_{crisis}$ *or* $suicidal_{ideation}$:

$show_{immediate_{crisis_{message(n)}}}$

$escalate_{to_{human_{clinician(n)}}}$

$\log_{event(\text{crisis_escalation})} n$)

continue # skip RL recommendation for this timestep

Step* 3. *Compute risk level

if $S_t.risk_{score} \geq S.risk_{\text{high}}$:

$action_{space} = \left\{low_{intensity_{interventions}}, clinician_{contact}, crisis_{protocol}\right\}$

else if $S_t.risk_{score} \geq S.risk_{medium}$:

$action_{space} = \left\{moderate_{interventions}, check_{in_{within_{short_{window}}}}\right\}$

else:

$action_{space} = \{educational_{content}, micro\text{-}interventions, gamified_{tasks}\}$

***Step* 4. *Policy decision* (*with simple if/else fallback*)**

if $random() < \varepsilon$:

$A_t = sample_{random(action_{space})}$

else:

$A_t = \pi\theta.select_{action(S_t,\ action_{space})}$

Step* 5. *Apply action / recommend content

$deliver_{action_{to_{user(A_t)}}}$

Step* 6. *Collect feedback & compute reward

$feedback = collect_{feedback(user_{response},\ engagement_{metrics},\ follow_{up_{assessments}})}$

$R_t = compute_reward(feedback, short_{term_{engagement_{weight}}}, long_{term_{outcome}}_weight)$

Step* 7. *Store transition and update personalization

$B.push\left(\left(S_t, A_t, R_t, S_{\{t+1\}}, feedback\right)\right)$

$update_{UserProfile(U,\ feedback,\ A_t)}$

***Step* 8. *RL update step* (*policy improvement*)**

if $enough_{samples_{in(B)}}$:

$batch = B.sample(minibatch_{size})$

$loss = compute_{RL_{loss(batch,\ \gamma)}}$

$\theta = \theta - \alpha * gradient(loss)$

Step* 9. *Post-delivery conditional checks for follow-up schedule

if feedback indicates $_{no_{engagement}}$ *or* $negative_{response}$:

if $U.recent_{rejections} >$ $threshold$:

$reduce_{recommendation_{frequency(U)}}$

$switch_{to_{lower_{dose_{interventions(U)}}}}$

else:

$schedule_{immediate_{follow_{up_{probe(n)}}}}$

else if feedback indicates $_{high_{engagement}}$ *and* $symptom_{improvement}$:

$reinforce_{similar_{actions_{in_{policy(\pi\theta,\ A_t)}}}}$

$schedule_{periodic_{booster(U)}}$

Step* 10. *Long-term evaluation & model maintenance

if $time_{to_{run_{periodic_{evaluation(n)}}}}$:

$evaluate_{policy\ on\ holdout_{users(n)}}$

if $performance_{drop_{detected}}$:

$retrain_{or\ finetune_{model(n)}}$

Step* 11. *Safety logging & human-in-loop alerting

if any $_{flagged_{events\ in(feedback,\ S_{\{t+1\}})}}$:

$notify_{clinical_{team(with_{context}=S_t,\ actions=A_t,\ feedback=feedback)}}$

end loop

The RL-AMHA algorithm 1 personalizes mental health education by observing user states, checking safety risks, and choosing interventions through reinforcement learning. If high risk, it triggers crisis protocols; otherwise, it recommends adaptive content. Feedback updates the policy, balancing short-term engagement and long-term well-being while maintaining safety and human-in-loop oversight. To specialize without training a distinct model for each user, RL-AMHA uses a learnt user embedding and LSTM-attention encoder to summarize each person's history. To reduce sparse feedback, the replay buffer mixes data from high and low-activity users, resulting in cohort-level gradients rather than a few heavy users. Multi-objective rewards handle conflicting feedback: high engagement but increased stress result in poor net returns due to well-being penalties, limiting dangerous behaviors. As more interactions occur, user embeddings and value assessments change from generic to personalized.

3.3 Recommendation & intervention engine

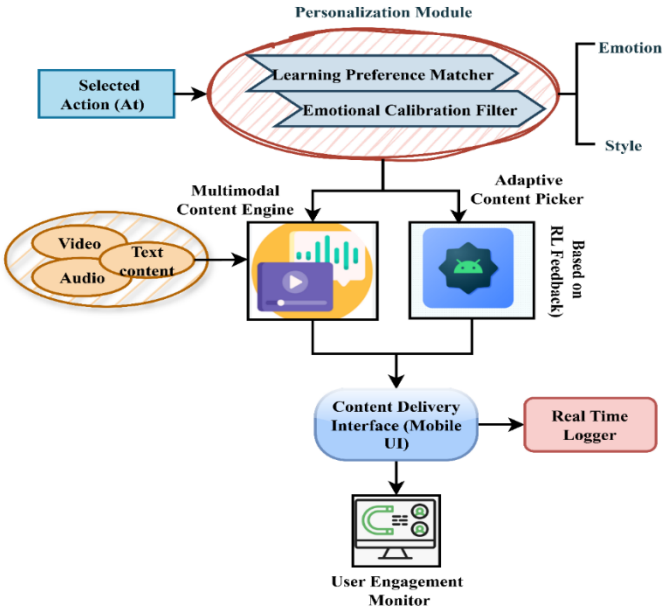


Figure 3: Recommendation & intervention engine

Figure 3 represents the personalized content delivery pipeline of RL-AMHA. The Selected Action (At) is processed by the personalization module, which synchronizes recommendations with the user's learning style and emotional state through the learning preference matcher and the emotional calibration filter. The output flows into the adaptive content picker, based on RL feedback, and the multimodal content engine, which can deliver video, audio, and text content. The final output is displayed using the content delivery interface, the user engagement monitor, and the real-time logger, which track various engagement metrics for learning and improvement.

Personalized content matching score N_u is expressed using equation 6,

$$N_u = \partial * MQN(b_u, v) \quad (6)$$

Equation 6 explains that the personalized content matching score, correspondence between the user's customized characteristics and the chosen action is measured by the matching score.

In this N_u is the composite matching score for action, $MQN(b_u, v)$ is the learning preference matcher score evaluating the suitability of the action for the user's learning style, ∂ is the tunable weighting parameter, b_u is the selected action at the time, and v is the target user.

Adaptive content selection probability $Q(d_j|b_u, v)$ is expressed using equation 7,

$$Q(d_j|b_u, v) = \frac{R(d_j, b_u, v)}{\theta} \quad (7)$$

Equation 7 explains the adaptive content selection probability given action and user, this soft-max-based equation calculates the likelihood of selecting a content item from a collection of candidate pieces.

In this $Q(d_j|b_u, v)$ is the probability of selecting content, $R(d_j, b_u, v)$ is the estimated quality score or expected reward of content, θ is the temperature parameter controlling randomness in selection, d_j is the specific content item among candidates, b_u is the selected action at the time, and v is the target user.

3.4 Feedback loop & reward estimation

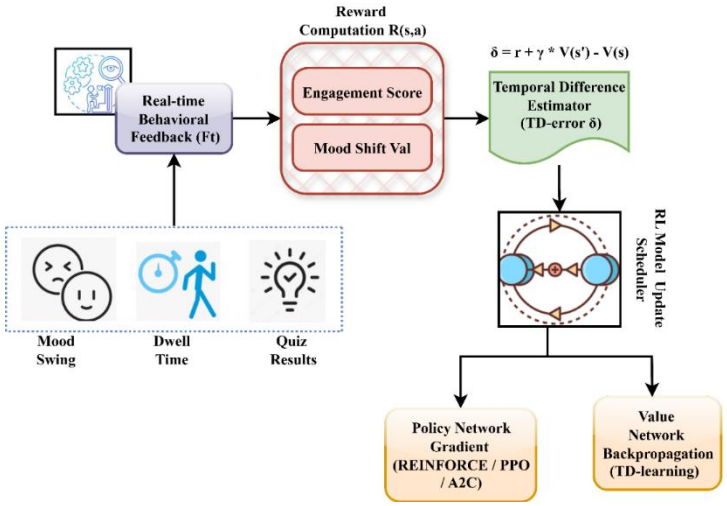


Figure 4: Feedback loop & reward estimation

Figure 4 illustrates RL-AMHA's feedback and learning loop. The real-time behavioral feedback (Ft), including mood changes, dwell time, and quiz performance, is used in Reward Computation R(s, a). The R(s, a) module processes engagement scores with the mood shift values. The Temporal Difference Estimator computes the TD-error (δ), a measure of prediction error. The RL Model Update Scheduler schedules fresh training updates (where gradients are computed) to the policy network (using either REINFORCE, PPO, or A2C) and to the Value Network (using TD-learning backpropagation). Thus, the RL Model Update scheduler is the final piece of our RL-AMHA feedback and learning loop, providing ongoing training and policy updates that refine the model's recommendations and improve the accuracy of its implications for a user's long-term mental health outcomes.

Reward computation from behavioral feedback $S(t_u, b_u)$ is expressed using equation 8,

$$S(t_u, b_u) = x_1 * F_u + x_2 * \nabla N_u \quad (8)$$

Equation 8 explains the computation of rewards from behavioral feedback, aggregating several engagement indicators, such as engagement score, mood change, and quiz performance.

In this $S(t_u, b_u)$ is the scalar reward signal for taking action, F_u is the real-time engagement score, ∇N_u is the mood shift value, representing change in emotional state after, and x_1, x_2 are the non-negative scalar weights balance components of the reward function.

Temporal difference error ∂_u is expressed using equation 9,

$$\partial_u = S(t_u, b_u) + \sigma W(t_{u+1}) \quad (9)$$

Equation 9 explains the temporal difference error between the expected value and the updated estimate, which combines the immediate payout and the discounted next-state value.

In this ∂_u is the temporal difference error at time, $S(t_u, b_u)$ is the immediate reward received for action, σ is the discount factor for future rewards, and $W(t_{u+1})$ is the estimated value of the next state.

Policy gradient update $\Delta_\tau K(\tau)$ is expressed using equation 10,

$$\Delta_\tau K(\tau) = F_\tau[\Delta_\tau * \mu_\tau(b_u|t_u) * \partial_u] \quad (10)$$

Equation 10 explains the policy gradient update expected return's gradient concerning policy parameters, decreasing slope in the log-probability.

In this $\Delta_\tau K(\tau)$ is the gradient of the objective function to policy parameters, $\mu_\tau(b_u|t_u)$ is the probability of action given state under a policy parameterized, ∂_u is the temporal difference error used as an advantage estimate, F_τ is the expectation over trajectories generated by policy, b_u is the action taken at time, t_u is the state at time, and τ is the policy network parameters.

3.5 System integration & continuous learning

Figure 5 illustrates RL-AMHA's systems integration and adaptive learning architecture. The mobile app/web portal UI interfaces with the recommendation API layer, providing personalized content. The frontend interactive module collects user interactions, and the real-time engagement tracker (RET) collects behavioral data. Data flows into cloud storage and the feedback + context repository, storing anonymized session details, histories, and reflections on mood shifts. The RL Training Scheduler operates through an online/offline update policy that employs batch replay memory and mini-batch fine-tuning. The model deployment pipeline utilizes a CI/CD policy that allows for canary and A/B testing methods to evaluate updates, enabling an automated, rolling deployment policy of updated policy $\pi(a|s)$ and value functions V(s).

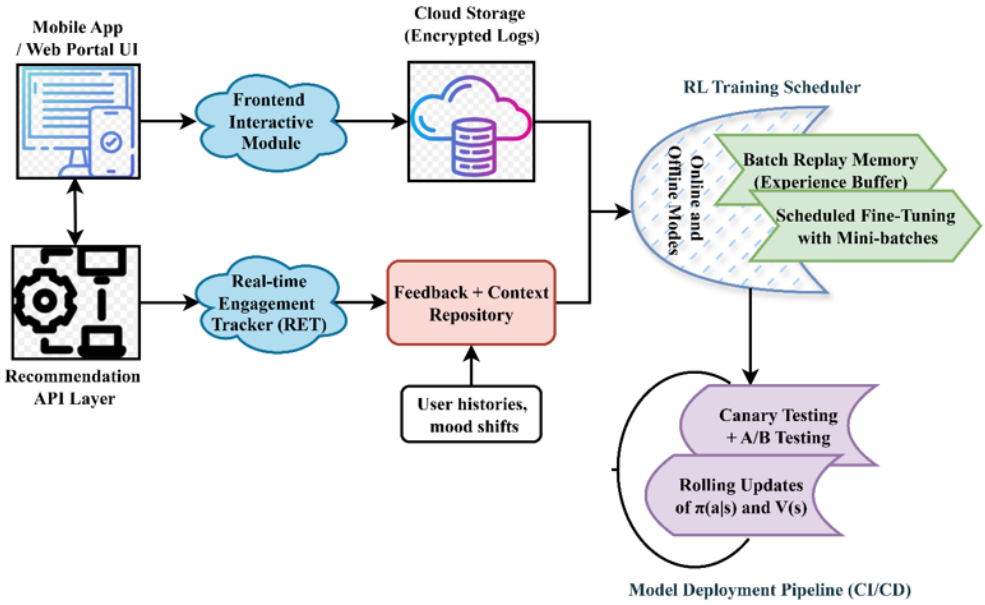


Figure 5: System integration & continuous learning

Batch replay memory update E_{u+1} is expressed using equation 11,

$$E_{u+1} = E_u \cup \{(t_u, b_u)\} \quad (11)$$

Equation 11 explains the batch replay memory update by adding the most recent experience tuple gathered at time, the rerun memory buffer is gradually updated.

In this E_u is the replay memory set at the iteration containing experience tuples, t_u is the state observed at time, b_u is the action taken at the time, and ∪ is the set union operator denoting the addition of new experience to memory.

Policy deployment with a/b testing probability $\mu^{dpy}(b|t)$ is expressed using equation 12,

$$\mu^{dpy}(b|t) = q * \mu^{nw}(b|t) + (1 - q) \quad (12)$$

Equation 12 explains the policy deployment policy combination of the new policy and the stable policy, currently, throughout canary testing.

In this $\mu^{dpy}(b|t)$ is the probability of an action given a state in the deployed policy, $\mu^{nw}(b|t)$ is the probability under the updated policy model, q is the mixing coefficient representing the fraction of users routed to the new policy, b is the candidate action, and t is the current user state.

A guarded updating loop controls continuous learning. Policy evaluations are conducted regularly on a fixed offline replay set and compared to a frozen reference policy. Any decrease in engagement, well-being, or safety measures prevents deployment. New models undergo canary/A-B tests on a limited user group and are reversed if online KPIs decline. To avoid forgetting, the replay buffer blends recent and older interactions, ensuring different historical states in mini-batches. Finally, mood, stress, and engagement drift indicators decrease update rates and require human assessment before large policy changes.

Evaluation metrics:

The evaluation metrics described utilize psychological, behavioral, and operational indicators to evaluate the RL-AMHA system's performance. The measures regarding satisfaction, engagement, stress reduction, frequency of activities taken up, barriers to implementation, resource utilization, and cultural fit provide a multi-factorial approach for measuring the effectiveness of the intervention and ensuring user-centered and context-relevant delivery of mental health education.

User satisfaction score T_u is expressed using equation 13,

$$T_u = \frac{1}{O} * t_j \quad (13)$$

Equation 13 explains that the user satisfaction score, the average rating provided by users, which represents

the perceived efficacy and quality of the service or material provided, is used to calculate.

In this T_u is the aggregate user satisfaction score, O is the total number of surveyed users, and t_j is the satisfaction rating from the user.

User engagement rate F_s is expressed using equation 14,

$$F_s = \frac{U_{ate}}{U_{ttl}} * 100\% \quad (14)$$

Equation 14 explains the user engagement rate as the proportion of users using interaction time relative to the total attainable session time.

In this F_s is the user engagement rate, U_{ate} is the duration of active user interaction within a session, and U_{ttl} is the total session duration available.

Stress level reduction ∇M_t is expressed using equation 15,

$$\nabla M_t = M_t^{pe} - M_t^{pt} \quad (15)$$

Equation 15 explains the stress level reduction: the difference, as determined by validated psychological measures.

In this ∇M_t is the magnitude of stress reduction, M_t^{pe} is the stress level before intervention, and M_t^{pt} is the stress level after intervention.

Engagement duration E_f is expressed using equation 16,

$$E_f = \frac{1}{N} * e_k \quad (16)$$

Equation 16 explains the engagement duration mean of each user session length over recorded sessions is used to calculate the average engagement time.

In this E_f is the mean engagement duration per session, N is the number of user sessions recorded, and e_k is the duration of the user session.

Self-care activity frequency G_{td} is expressed using equation 17,

$$G_{td} = \frac{B_x}{X} \quad (17)$$

Equation 17 explains that the average frequency of self-care actions completed by each user over a time window is measured by the self-care frequency.

In this G_{td} is the average self-care activity frequency per user per time window, B_x is the total of self-care activities recorded during the window, and X is the length of the time window.

Barriers to implementation index C_j is expressed using equation 18,

$$C_j = c_l * x_l \quad (18)$$

Equation 18 explains that the barriers-to-implementation index's total difficulty level is quantified as the weighted sum of the identified implementation-related hurdles.

In this C_j is the composite barrier to implementation score, c_l is the presence or magnitude of the barrier, and x_l is the weight representing the importance or impact of barrier.

Resource allocation efficiency S_f is expressed using equation 19,

$$S_f = \frac{V_s}{D_s} \quad (19)$$

Equation 19 explains that operational efficiency, which is indicated by the ratio of resources used that directly contribute to intervention efficacy over all resources used, is a measure of resource allocation efficiency.

In this S_f is the efficiency ratio for resource allocation, V_s is the amount of resources effectively used, and D_s is the total resources allocated to the project.

Cultural acceptance score D_b is expressed using equation 20,

$$D_b = \frac{1}{O_d} * d_n \quad (20)$$

Equation 20 explains the cultural acceptance score measure of how well the program conforms to social and cultural standards, which is averaged from the community of user feedback evaluations.

In this D_b is the mean cultural acceptance rating, O_d is the number of respondents evaluating cultural compatibility, and d_n is the respondent's cultural acceptance rating.

The equation 21, FTJ calculates emotional state improvement score that have positive sign.

$$FTJ = [A_u - A_{tu} - A_{boz} + A_i] \quad (21)$$

Linear combinations are adjusted A_u and scaled A_{tu}, and then clamped using the clamp function. The index can gauge balanced changes in multiple affective dimensions A_{boz} while minimizing harm A_i from outlier distortions.

Equation 22 specifies the behavioral engagement index CFJ as a saturating exponential utility, with adherence components, content completion.

$$CFJ = \left[1 - f^{-(\forall d + \propto x + \tau\epsilon)}\right] * (1 - r) \quad (22)$$

A multiplicative hazard penalty ($\forall d$) reduces the overall index $\propto x$ by accounting for the churn $\tau\epsilon$, or the probability of disengagement r.

Together, the metrics reflect how well the user is doing as well as the efficiency of the system. They allow for continuous improvement by capitalizing on successes, identifying areas for improvement, and ensuring that users' cultural and resource considerations align with the provision of mental health recommendations. Holistically, these metrics enable scalable, flexible, and actionable mental health recommendations to optimize users' long-term psychological well-being.

Learning a closed-loop policy that optimizes intervention sequences, not just single-step predictions, makes RL-AMHA better than other adaptive approaches. Unlike supervised neural networks or fuzzy controllers, it updates online from each interaction, investigates alternative tactics, and improves engagement, stress reduction, and self-care via a multi-objective reward. This makes behavioral drift adaption and engagement recovery faster than batch-retrained neural recommenders or manually modified fuzzy rules. While output-feedback controllers presume simple dynamics

and objectives, RL-AMHA's LSTM-attention backbone handles high-dimensional, partially seen states for richer temporal modeling and data-driven customisation.

4 Results and discussion

This section presents an assessment of resource utilization, contextual preferences, and RL-AMHA's performance compared to other materials, while also examining how reinforcement learning optimization enhanced personalization, user engagement, and psychological outcomes. It presents evidence of RL-AMHA's superior performance and discusses its ability to provide scalable, effective mental health education.

4.1 Dataset description

The student mental health & resilience dataset on kaggle contains behavioral, emotional, and academic data with stress labels available. The factors it can tap into include psychometric indicators such as mood, stress markers, engagement, and academic behavior and performance among students [24]. This multimodal dataset is conducive to modeling the dynamics of mental states and resilience factors. It enables the training and testing of learning platforms that utilize reinforcement learning, such as RL-AMHA. The semi-complex scans enable the exploration of personalized intervention strategies based on students' real-life behavior.

The Kaggle "Student Mental Health & Resilience" dataset was cleaned, one-hot encoded, and z-normalized before being divided by student into 70% train, 15% validation, and 15% test partitions. Over time, student records were sorted into RL episodes. In pretraining, stress, anxiety, depression-risk, and resilience categories were mapped to 0-1 scores for use in the state vector and as supplementary supervised targets. The multi-objective signal used by RL-AMHA was created by combining changes in well-being (lower stress/higher mood), engagement (more participation and stable grades), self-care behavior, and penalties for deterioration or drop-out, normallyized to [-1, 1] and linearly weighted.

Table 2: Student mental health & resilience dataset

Attribute	Description
Dataset Name	Student Mental Health & Resilience Dataset
Source	Kaggle – Ziya07 (Link)
Content Type	Behavioral, emotional, and academic indicators with stress-related labels
Key Features	Mood scores, stress levels, engagement metrics, and academic performance
Purpose	Modeling mental state dynamics and resilience factors
Use Case	Training/evaluating reinforcement learning–based adaptive systems (RL-AMHA)
Application	Development of personalized mental health intervention strategies

4.2 User satisfaction (Score 1–5)

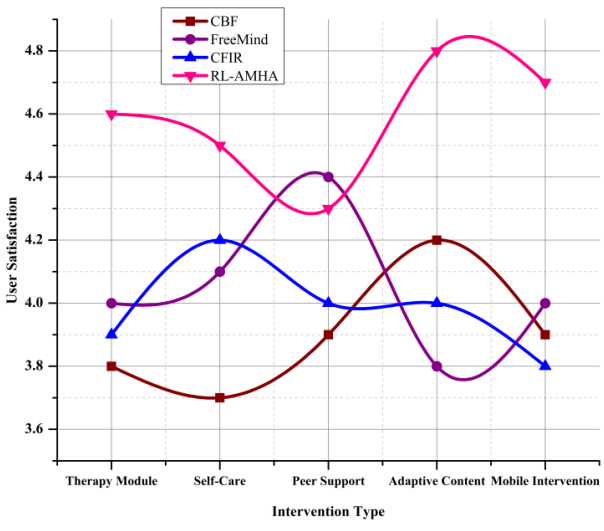


Figure 6: User Satisfaction (Score 1–5)

Figure 6 illustrates user satisfaction across the five types of interventions - Therapy Module, Self-Care, Peer Support, Adaptive Content, and Mobile Intervention - for four systems: CBF, FreeMind, CFIR, and RL-AMHA. RL-AMHA is leading the other systems made evaluated using equation 13. It receives the highest overall score, which is around 4.85, in adaptive content. FreeMind is rated moderately, with a rating of approximately 4.0 to 4.4, and CFIR is also rated at peak, with Self-Care receiving a rating of approximately 4.2. However, all of these ratings are fluctuating, and CBF has been rated the lowest throughout the entire process, with a rating of approximately 3.55 to 4.0. The results of this study reveal that RL-AMHA is clearly superior in terms of adaptation and personalization across a wide range of mental health interventions, particularly in the field's mobile-based and adaptive modules.

4.3 User engagement (%)

Figure 7 shows user engagement (%) across five weeks for four systems: CBF, FreeMind, CFIR, and RL-AMHA. In user engagement, RL-AMHA measures steady user engagement and achieves the highest usage, engaging all users made evaluated using equation 14.

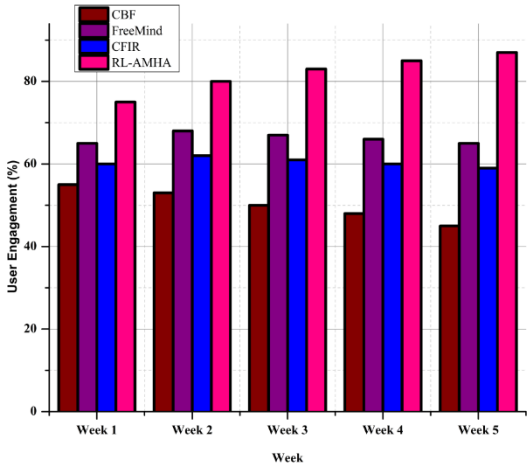


Figure 7: User engagement (%)

User engagement in RL-AMHA increased from ~75% in week 1 to ~88% in week 5, suggesting high retention and responsiveness. FreeMind showed slight variation and relatively high user engagement stats as well (~65 increase - ~68 in Week 5). CFIR shows a moderate level of user engagement (~59 - ~62), with slight fluctuations. CBF scored lowest as user engagement declined from about ~55% to ~46% over time. Again, the key takeaway is that RL-AMHA demonstrates its ability to promote user participation over extended periods of use.

4.4 Stress level reduction (score)

Figure 8 shows the reduction in stress levels from pre-intervention to post-intervention week 4 (five time points) for four systems: RL-AMHA, CFIR, FreeMind, and CBF. All four systems started at similar reductions in stress level (7-7.8), although each method had a different trajectory for lowering stress level made evaluated using equation 15. Throughout the course, observe RL-AMHA consistently delivering high performance, with gradual improvements evident post-week 4. See CFIR and FreeMind with moderate, stable reductions. CBF remains relatively robust despite the decrease in stress levels, although with less improvement than the other methods. The data show that RL-AMHA has a longer-term impact on stress reduction and can sustain its progress over time compared to other methods.

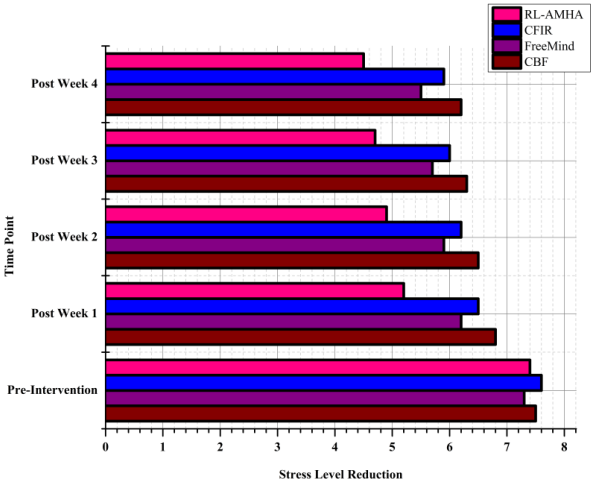


Figure 8: Stress level reduction (score)

4.5 Engagement duration (minutes)

Figure 9 shows the total engagement time by users during five sessions for each system, CBF, FreeMind, CFIR, and RL-AMHA. The following overview captures these graphs. For each method, users continued to show more engagement time, except for CBF, which shows engagement time declining from ~15 to ~12 minutes made evaluated using equation 16. RL-AMHA delivered the longest engagement time; it grew from ~22 to nearly 29 minutes by Session 5. Both FreeMind and CFIR show milder upward trends, with near peaks of engagement time around 21 and 19. Overall, these findings depict that RL-RL-AMHA showed the most consistent increase in user Engagement from Session 1 to Session 5.

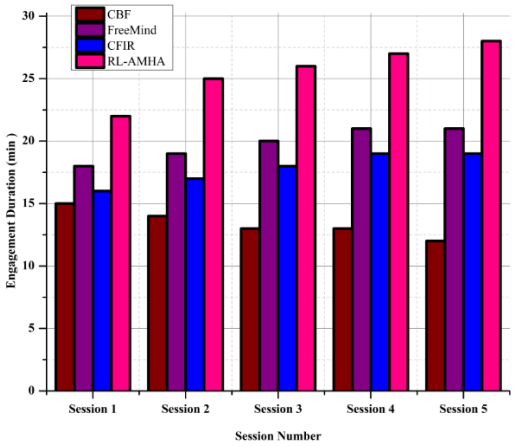


Figure 9: Engagement duration (minutes)

4.6 Self-care activity frequency (count)

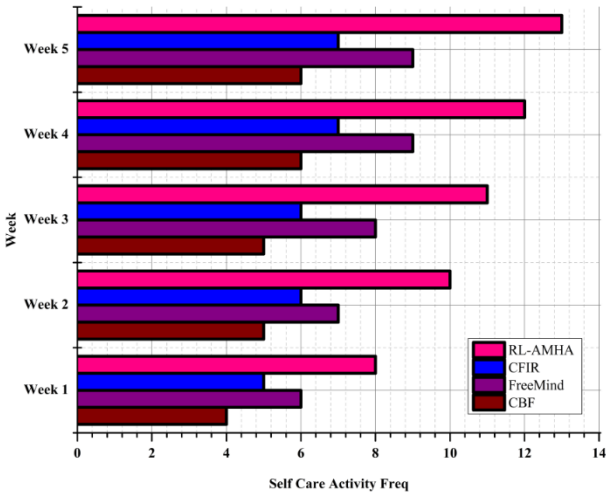


Figure 10: Self-care activity frequency (count)

Figure 10 illustrates the frequency of self-care activities over 5 weeks for RL-AMHA, CFIR, FreeMind, and CBF. RL-AMHA is best, increasing from ~10 activities in week 1 to ~13 in week 5. CFIR shows steady improvement throughout the five weeks, peaking at below ~11 activities. FreeMind showed a modest increase, around ~9 by week 5, while CBF showed tiny movement, from ~4 to ~7 activities, evaluated using equation 17. Overall, this evidence highlights RL-AMHA's ability to encourage ongoing engagement and increase self-care engagement compared to a more traditional baseline and alternative self-care activity systems.

4.7 Barriers to implementation (% frequency)

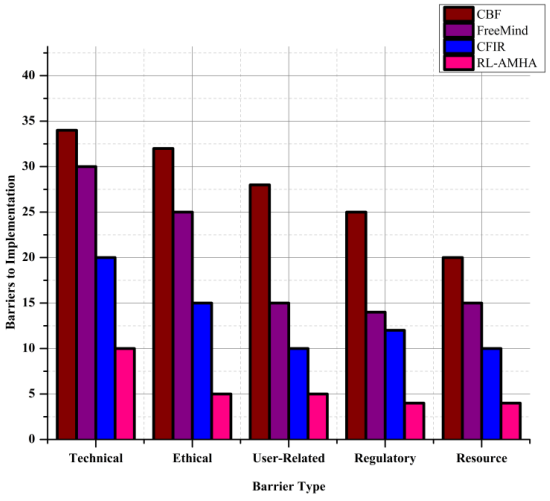


Figure 11: Barriers to implementation (% frequency)

Figure 11 shows a visual comparison of barriers to implementation for CBF, FreeMind, CFIR, and RL-AMHA across five categories: Technical, Ethical, User-Related, Regulatory, and Resource. CBF appears to have the most barriers consistently across categories, especially Technical (~34) and Ethical (~32) evaluated using equation 18. FreeMind and CFIR appear to face the second-highest challenges, and encompass both technical and ethical barriers. RL-AMHA across categories exhibited the lowest levels of implementation obstacles, with apparent reductions in both technical (~10) and moral (~5) barriers. This ultimately demonstrates that RL-AMHA offers the

best adaptability, the simplest implementation, and fewer barriers than the other methods.

4.8 Resource Allocation (%)

Table 3: Resource allocation (%)

Resource Type	CBF	FreeMind	CFIR	RL-AMHA
Time	30	35	40	50
Personnel	25	30	35	45
Funding	20	25	30	40
Technology	15	20	25	35
Training	10	15	20	30

Table 3 illustrates resource usage for CBF, FreeMind, CFIR, and RL-AMHA across five resource types: Time, People, Money, Technology, and Training. RL-AMHA consistently requires the most resources across all models, with time (50) and people (45) substantially higher than everything else, which is fitting given that RL-AMHA is the most advanced in terms of computational complexity and operational efficiency, as evaluated using equation 19. CFIR is second, and then appears FreeMind, and CBF uses the least resources of the four options. In sum, the gradient seems to signal a trade-off between increasing performance and total resource use (time, people, money, technology, & training) for the more advanced, AI-based approaches (RL-AMHA).

4.9 Cultural acceptance (%)

Table 4: Cultural Acceptance (%)

Demographic Group	CBF	FreeMind	CFIR	RL-AMHA
Group A	65	70	75	85
Group B	60	68	72	83
Group C	58	65	70	80
Group D	55	63	68	78
Group E	50	60	65	75

Table 4 data set shows the effectiveness or penetration rates of CBF, FreeMind, CFIR, and RL-AMHA across five groups (A-E); RL-AMHA consistently had the highest values, with a peak of 85 in Group A and remained the highest across all demographics. CFIR ranks second, followed by FreeMind and CBF. There is a gap between RL-AMHA and traditional methods. The gap is larger for higher-performing groups evaluated using equation 20. This suggests RL-AMHA is more easily adaptable and has a broader appeal, with implications that it offers greater personalization and context-specific adaptation, leading to higher acceptance across the populace.

4.10 Emotional state improvement score and behavioral engagement index

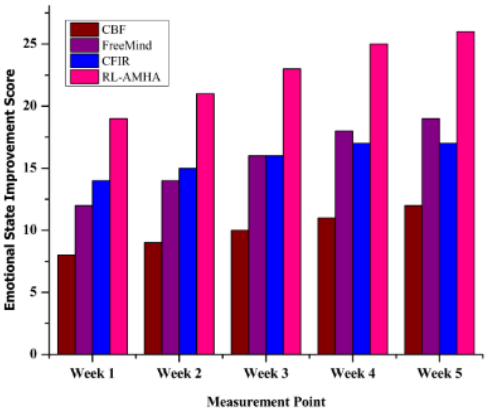


Figure 12(a): Emotional State Improvement Score

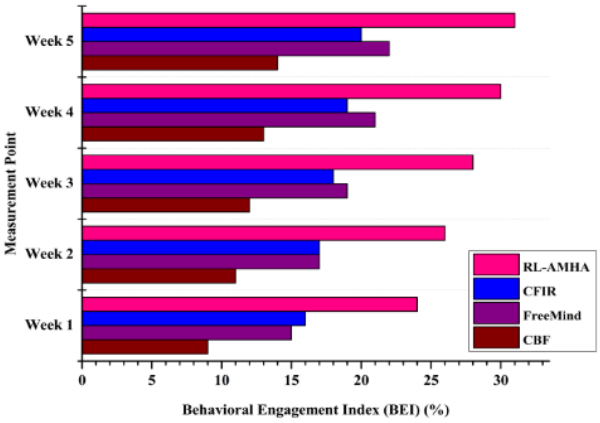


Figure 12(b): Behavioural Engagement Index

Figure 12(a) illustrates the Emotional State Improvement Score (ESIS), which tracks the percentage change in emotional well-being from pre- to post-intervention, based on surveys, sentiment analysis, or mood tracking via wearables. Figure 12(b) illustrates the Behavioral Engagement Index (BEI), which measures positive mental health behaviors comprising mindfulness practice, support for activity engagement, and adherence to coping strategies. Both ESIS and BEI provide a synthesized assessment of the emotional and behavioral consequences of the personalized recommendations, including the extent to which an intervention improves mood and increases long-term, healthy attitudes towards mental health maintenance. The user satisfaction across the five types of interventions of therapy module, self-care, peer support, adaptive content, and mobile intervention, for four systems. It includes the CBF, FreeMind, CFIR, and RL-AMHA. RL-AMHA is leading the other systems evaluated using Equation 21. It scores the highest overall at ~4.85 in adaptive content on FreeMind is rated moderately ~4.0 - 4.4 as well as CFIR - at peak, Self-Care ~4.2 too, but is all in fluctuation, and CBF is rated lowest all along (~3.55 - 4.0) and evaluated using equation 22. These findings demonstrate RL-AMHA's clear superiority in adaptability and personalization across a range of mental health interventions, especially in the field's adaptive and mobile-based modules. The findings of this evaluation confirm that RL-AMHA outperformed CBF, FreeMind, and CFIR on all evaluation metrics. It was less resource-intensive, could be adapted to larger and more diverse demographic communities, and provided better recommendations. All of which highlight the success of RL-AMHA in personalizing mental health education, a scalable, data-driven tool for improving psychosocial outcomes. Besides charting, the Behavioral Engagement Index (BEI) predicts retention and is a reward word. The hazard-penalty function devalues short, inconsistent sessions and rewards frequent, stable engagement, resulting in a single scalar summary of use intensity and consistency. RL-AMHA's higher BEI scores result in higher 5-week retention probabilities than CBF, FreeMind, and CFIR, indicating that optimizing BEI improves user retention outcomes. The reinforcement learning system has three key hyperparameters: exploration rate (ε), discount factor (γ), and learning rate (α). It took three steps to tune these. First, a coarse grid search spanning large ranges (ε: 0.05–0.40, γ: 0.85–0.99, α: 1e-5–1e-2) found stable regions. Second, Bayesian optimization with offline simulations under safety constraints resulted in ε≈0.18, γ≈0.95, and $\alpha \approx 2\times10^{-4}$. An online safety-adaptive adjustment was implemented, with ε decreasing over time but never below 0.05, α adjusting based on reward trends, and γ remaining constant to balance short- and long-term wellbeing. Replay testing, clinician assessment, and controlled A/B trials confirmed steady learning and no harmful recommendations. Ablation experiments trained four PPO variants: full reward, engagement-only, well-being-only, and no-penalty. In the whole model, engagement averaged 88%, session duration 29.0 minutes, stress reduction gain 18.6%, and self-care frequency 13.1 actions/week. Although engagement increased session time to 31 minutes (+6-8%), it decreased stress improvement to ≈14% and self-care to ≈11/week. Well-being improved stress by 19% but decreased involvement to 83% and self-care to 10/week. Removing penalties enhanced early engagement but increased mid-study dropout. Compared to all ablated alternatives, the full reward significantly improved engagement and well-being ($p < 0.05$).

5 Discussion

To represent mental health support, RL-AMHA uses an LSTM-attention encoder to store mood, behavior, and intervention history sequentially, rather than treating each encounter separately, as in CBF or hybrid recommenders. Improved state representation enables PPO to modify policies to user context, leading to increased satisfaction with therapy, self-care, peer

support, adaptive content, and mobile interventions. RL-AMHA optimizes a multi-objective reward that combines engagement, stress reduction, and self-care adherence, leading to longer session durations and increased activity frequency over weeks, unlike CFIR and implementation-only frameworks. Static recommenders and rule-based mobile interventions cannot recover from user disengagement or stress changes, resulting in plateaued or declining satisfaction and engagement. This is seen in CBF's shorter sessions and lower weekly involvement and FreeMind/CFIR's slower activity frequency and emotional improvement. RL-AMHA adjusts its policy based on interaction rewards, correcting prior decisions and emulating successful patterns, resulting in compounding benefits over weeks. A closed-loop RL design makes RL-AMHA ideal for real-time feedback from mobile and sensor streams, allowing policy revisions without rebuilding the intervention program. PPO policy fine-tuning allows practitioners to prioritize objectives, such as stress relief for clinical users or preventive engagement, while maintaining LSTM focus. Given its flexibility and low implementation barriers, RL-AMHA can be used as an adaptive personalization layer on top of existing MIBIs, self-care apps, or peer-support platforms, rather than a full system replacement. RL-AMHA is a data-driven variant of adaptive and robust mental health control. Similar to adaptive fuzzy, neural adaptive, output-feedback, backstepping, and nonlinear optimum control, it addresses uncertainty and optimizes long-term results. RL-AMHA learns its policy from interaction data using LSTM-based state representation and multi-objective reward, eliminating the need for explicit models or rules. This makes it more scalable to high-dimensional, partially witnessed behavioral data and easier to adjust to changing objectives. RL-AMHA maintains optimal control and adaptive adjustment for long-term goals without making significant assumptions about human psychology. RL-AMHA protects privacy and ethics by design. All interaction data are collected only with informed consent, stored in encrypted form, and de-identified before model training; access is restricted and only aggregate metrics are used in monitoring. The system is positioned as a low-risk support tool, not a diagnostic or emergency service, and high-risk patterns trigger escalation to human professionals rather than autonomous action. The reward function penalizes strategies that increase stress or exploit engagement in unhealthy ways, discouraging “addictive” behaviors. Users retain control through opt-out and deletion options, and recommendations are explained in plain language to support transparency and trust.

6 Conclusion and future work

RL-AMHA, which provides highly individualized interventions and mental health education. Comparisons between CBF, FreeMind, and CFIR demonstrated that RL-AMHA can reduce both cost and time, provide individualized recommendations for diverse demographics, and offer accurate insights. This model can leverage reinforcement learning to continuously optimize its content delivery based on end-user feedback, thereby promoting ongoing engagement and improving psychological outcomes.

In the future, this framework can be adapted to include multiple categories of multimodal data sources, such as speech, facial expressions, and physiological signals, thereby deepening understanding of emotion and behavior. Integration of wearable devices and mobile health platforms would enable real-time intervention. Scaling the system for multilingual and cross-cultural contexts would likewise reach a global audience. Building explainable AI into the intervention approach will increase trust and transparency, making RL-AMHA a holistic, ethical, and scalable solution for mental health education and support.

Funding

Project of Henan Province's Philosophy and Social Sciences for Building an Education-Strong Province: The Influence of Childhood Emotional Trauma on Adolescents' Emotional Stability and Intervention - Based on the Theory of Cognitive Reappraisal (No. 2025JYQS0318)

Competing interests

The author declared no potential conflicts of interest with respect to the research, authorship or publication of this article.

Author contributions

Wang. Project administration, Conceptualization, Research concept and design&Data analysis and visualization&Review and revision

Data availability statement

All data generated or analysed during this study are included in this article.

References

[1] Saleem, M., Kühne, L., De Santis, K. K., Christianson, L., Brand, T., & Busse, H. (2021). Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR mental health, 8(12), e30000. DOI: 10.2196/30000

[2] Balaskas, A., Schueller, S. M., Cox, A. L., & Doherty, G. (2021). Ecological momentary interventions for mental health: A scoping review. PloS one, 16(3), e0248152. DOI: 10.1371/journal.pone.0248152

[3] Ahad, A. A., Sanchez-Gonzalez, M., & Junquera, P. (2023). Understanding and addressing mental health stigma across cultures for improving psychiatric care: A narrative review. Cureus, 15(5). doi: 10.7759/cureus.39549

[4] Buselli, R., Corsi, M., Veltri, A., Baldanzi, S., Chiumiento, M., Del Lupo, E., ... & Cristaudo, A. (2021). Mental health of Health Care Workers (HCWs): a review of organizational interventions put in place by local institutions to cope with new psychosocial c

hallenges resulting from COVID-19. Psychiatry res earch, 299, 113847. DOI: 10.1016/j.psychres.2021. 113847

[5] Thornicroft, G., Sunkel, C., Aliev, A. A., Baker, S., Brohan, E., El Chammay, R., ... & Winkler, P. (202 2). The Lancet Commission on ending stigma and d iscrimination in mental health. The Lancet, 400(103 61), 1438-1480. doi: 10.1016/S0140-6736(22)0147 0-2.

[6] Fleming, W. J. (2024). Employee well-being outco mes from individual-level mental health interventio ns: Cross-sectional evidence from the United Kingd om. Industrial Relations Journal, 55(2), 162-182. ht tps://doi.org/10.1111/irj.12418

[7] Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl , S. E., Vargas, J., Minkel, J., ... & Zilca, R. (2021). Artificially intelligent chatbots in digital mental hea lth interventions: a review. Expert review of medica l devices, 18(sup1), 37-49. DOI: 10.1080/17434440 .2021.2013200

[8] Fusar-Poli, P., Correll, C. U., Arango, C., Berk, M., Patel, V., & Ioannidis, J. P. (2021). Preventive psyc hiatry: a blueprint for improving the mental health o f young people. World Psychiatry, 20(2), 200-221. DOI: 10.1002/wps.20869

[9] Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y. H., ... & Torous, J. (2025). Large language models i n mental health care: a scoping review. Current Tre atment Options in Psychiatry, 12(1), 1-18. https://do i.org/10.1007/s40501-025-00363-y

[10] Borghouts, J., Eikey, E., Mark, G., De Leon, C., Sc hueller, S. M., Schneider, M., ... & Sorkin, D. H. (2 021). Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. Journal of medical Internet research, 23(3), e24387. DOI: 10.2196/24387

[11] Merino, M., Tornero-Aguilera, J. F., Rubio-Zarapuz , A., Villanueva-Tobaldo, C. V., Martín-Rodríguez, A., & Clemente-Suárez, V. J. (2024, July). Body pe rceptions and psychological well -being: a review of the impact of social media and physical measureme nts on self-esteem and mental health with a focus on body image satisfaction and its relationship with cu ltural and gender factors. In Healthcare (Vol. 12, No . 14, p. 1396). MDPI. https://doi.org/10.3390/health care12141396

[12] Lattie, E. G., Stiles-Shields, C., & Graham, A. K. (2 022). An overview of and recommendations for mor e accessible digital mental health services. Nature R eviews Psychology, 1(2), 87-100. https://doi.org/10. 1038/s44159-021-00003-1

[13] D. Zaçe, I. Hoxhaj, A. Orfino, A. M. Viteritti, L. Ja niri, and M. L. Di Pietro, “Interventions to address mental health issues in healthcare workers during in fectious disease outbreaks: A systematic review,” Jo urnal of Psychiatric Research, vol. 136, pp. 319–33 3, Apr. 2021, doi: https://doi.org/10.1016/j.jpsychire s.2021.02.019.

[14] D. Z. Q. Gan, L. McGillivray, J. Han, H. Christense n, and M. Torok, “Effect of Engagement With Digit al Interventions on Mental Health Outcomes: A Sys tematic Review and Meta-Analysis,” Frontiers in Di gital Health, vol. 3, Nov. 2021, doi: https://doi.org/1 0.3389/fdgth.2021.764079.

[15] T. J. Philippe et al., “Digital Health Interventions fo r Delivery of Mental Health Care: Systematic and C omprehensive Meta-Review (Preprint),” JMIR Men tal Health, vol. 9, no. 5, 2021, doi: https://doi.org/10 .2196/35159.

[16] Mazlan, I., Abdullah, N., & Ahmad, N. (2023). Exp loring the impact of hybrid recommender systems o n personalized mental health recommendations. Inte rnational Journal of Advanced Computer Science an d Applications, 14(6), 935-944. DOI: 10.14569/IJA CSA.2023.0140699

[17] [Mazlan, I. B., Abdullah, N., Ahmad, N., & Harun, S. Z. (2024). A Review of Personalized Recommen der System for Mental Health Interventions. Interna tional Journal of Advanced Computer Science & Ap plications, 15(10). DOI: 10.14569/IJACSA.2024.01 51041

[18] Kazi, F. F., & Sandbulte, J. (2023). Examining a soc ial-based system with personalized recommendation s to promote mental health for college students. Sm art health, 28, 100385. DOI: 10.1016/j.smhl.2023. 100385

[19] Singh, V., Kumar, A., & Gupta, S. (2022). Mental h ealth prevention and promotion—A narrative revie w. Frontiers in psychiatry, 13, 898009. https://doi.or g/10.3389/fpsyt.2022.898009

[20] Worsley, J. D., Pennington, A., & Corcoran, R. (20 22). Supporting mental health and wellbeing of univ ersity and college students: A systematic review of r eview-level evidence of interventions. PLoS one, 17 (7), e0266725. DOI: 10.1371/journal.pone.0266725

[21] Hornstein, S., Zantvoort, K., Lueken, U., Funk, B., & Hilbert, K. (2023). Personalization strategies in di gital mental health interventions: a systematic revie w and conceptual framework for depressive sympto ms. Frontiers in digital health, 5, 1170002. DOI: 10. 3389/fdgth.2023.1170002

[22] Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxho orn, J., Jones, P. B., Patalay, P., ... & Griffiths, S. L. (2024). The social determinants of mental health an d disorder: evidence, prevention and recommendati ons. World psychiatry, 23(1), 58-90. DOI: 10.1002/ wps.21160

[23] Mutschler, C., Bellamy, C., Davidson, L., Lichtenst ein, S., & Kidd, S. (2022). Implementation of peer s upport in mental health services: A systematic revie w of the literature. Psychological Services, 19(2), 3 60. DOI: 10.1037/ser0000531

[24] https://www.kaggle.com/datasets/ziya07/student-mental-health-and-resilience-dataset

[25] Kolenik, T., & Schiepek, G. (2024). Computational psychotherapy system for mental health prediction and behavior change with a conversational agent. Neuropsychiatric Disease and Treatment, 20, 2465–2498. https://doi.org/10.2147/NDT.S417695

[26] Kolenik, T. (2025). Intelligent cognitive system for computational psychotherapy with a conversational agent for attitude and behavior change in stress, anxiety, and depression. Informatica, 49(2), 259–272. https://doi.org/10.31449/inf.v49i2.8738

[27] Kolenik, T. (2021). Methods in digital mental health: Smartphone-based assessment and intervention for stress, anxiety and depression. In C. Comito, A. Forestiero, & E. Zumpano (Eds.), Integrating artificial intelligence and IoT for advanced health informatics (pp. xx–xx). Springer. https://doi.org/10.1007/978-3-030-91181-2_7

[28] Kolenik, T., & Gams, M. (2021). Intelligent cognitive assistants for attitude and behavior change support in mental health: State-of-the-art technical review. Electronics, 10(11), 1250. https://doi.org/10.3390/electronics10111250

[29] Kolenik, T., & Gams, M. (2021). Persuasive technology for mental health: One step closer to (mental health care) equality? IEEE Technology and Society Magazine, 40(1), 80–86. https://doi.org/10.1109/MTS.2021.3056288

[30] Boulkroune, A., Zouari, F., & Boubellouta, A. (2025). Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control. Advance online publication. https://doi.org/10.1177/10775463251320258

[31] Boukabou, A. (2017). Output-feedback controller-based projective lag-synchronization of uncertain chaotic systems in the presence of input nonlinearities. Mathematical Problems in Engineering, 2017, Article 8045803. https://doi.org/10.1155/2017/8045803.

[32] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. Energy Reports, 9, 1470–1486. https://doi.org/10.1016/j.egyr.2023.03.020

H-CMAF: A Deep Learning-Based Cross-Modal Attention Fusion Framework for Dangerous Goods Detection from X-Ray and Visible Light Imagery

Haojie Ren
Shanxi Police College, Taiyuan, Shanxi, 030401, China
E-mail: hj-ren-best@outlook.com

Keywords: multimodal fusion, deep learning, dangerous object detection, cross-modal attention mechanism, smart policing

Recieved: October 13, 2025

Hazardous goods identification still faces challenges such as low accuracy and difficulty identifying items in low-light, densely packed environments. Therefore, this paper combines multimodal fusion information with deep learning models to construct an identification and classification system. This aims to improve the accuracy of hazardous goods identification in the public domain and its robustness in diverse environments and conditions. Based on the attention mechanism, this study fuses feature information from X-rays and visible light to construct an H-CMAF framework model. In challenging scenarios such as low light, partial occlusion, and dense stacking, H-CMAF achieves minimal mAP degradation (6.5%, 5.9%, and 7.1%, respectively), demonstrating exceptional robustness. Furthermore, the model achieves an inference speed of 22.8 FPS on edge devices, with only 34.8M parameters and a memory footprint of 100MB, outperforming most competing models and achieving a good balance between accuracy and efficiency.

Povzetek: Prispevek predstavi učinkovit večmodalni (rentgen + vidna svetloba) model z mehanizmom pozornosti za robustnejšo in natančnejšo identifikacijo nevarnih predmetov tudi v zahtevnih pogojih (slaba svetloba, zakritost, gneča) ter z dobro hitrostjo delovanja na robnih napravah.

1 Introduction

As the world enters the information age, the intelligentization of science and technology has also brought challenges to public security. Terrorists and criminals have become more intelligent in their methods of committing crimes, and the frequency of large-scale events associated with public crises has increased dramatically. The control of dangerous goods has become an important issue in maintaining national security and public safety. In the field of public security, dangerous goods are not limited to highly offensive items such as firearms and ammunition, but also include new 3D-printed weapons represented by intelligent ones, non-metallic contraband, and other diverse forms. These items, due to their concealment, have caused tremendous pressure in the field of public security.

Most current research focuses on improving the accuracy of dangerous object recognition, but fails to consider the impact of complex real-world scenarios, such as occlusion and overlap, on accuracy. This paper aims to address this issue.

When studying the synchronization control strategies of uncertain nonlinear systems, various methods have been proposed to deal with different types of chaotic systems. For example, Boulkroune et al. [1] explored the practical finite-time fuzzy synchronization problem of fractional-order chaotic systems and proposed two chatter-free methods. Similarly, Bowong et al. [2] studied the synchronization and duration problems of a class of uncertain chaotic systems, providing a theoretical basis for understanding the dynamic behavior of chaotic systems. In addition, for the practical fixed-time synchronization of fractional-order chaotic systems, Rigatos et al. [3] further developed an adaptive fuzzy control strategy. For the robust control of uncertain nonlinear systems, the work of Yuan et al. [4] and others demonstrated how to use neural networks to achieve effective control of nonlinear systems with actuator amplitude and rate saturation. Finally, in the context of single-input single-output (SISO) nonlinear systems, Practical et al. [5] achieved uncertainty handling by adopting an adaptive backstepping method. These studies have jointly promoted the development of complex system control theory and provided valuable guidance for engineering practice.

In the field of hazardous materials inspection, it is impossible to fully detect safety hazards through video or X-ray alone. Therefore, a new direction is to fuse multimodal data and infer the classification results of hazardous materials through the complementary nature of multimodal information [6] .

Currently, the field of hazardous materials inspection has shifted from traditional image analysis methods to a data-driven multimodal approach. Most current deep learning research uses X-ray or visible light image analysis to identify hazardous materials. This study will combine multimodal information image classification and detection with deep learning models to build a framework specifically for hazardous materials detection. This paper builds a dedicated hazardous materials recognition model based on a series of methods. The proposed model will be compared with traditional single-modal methods and the latest deep learning models [7].

This study aims to address the information fusion challenge in multimodal hazardous materials detection. Specific objectives are as follows: (1) Design a robust cross-modal fusion framework to cope with occlusion and illumination changes in a single modality; (2) Explore the role of dynamic attention mechanisms in improving detection accuracy. To this end, we propose the H-CMAF model and use mAP, precision, recall, and FPS as core evaluation metrics.

The innovation of this paper lies in the first application of a dynamic cross-modal attention mechanism to dangerous goods detection, optimizing fusion weights specifically for occlusion scenarios, and designing a lightweight architecture suitable for deployment in real-world security equipment.

2 Literature review

2.1 X -ray

X-rays are widely used in security inspections because they can identify the internal structure of objects. Early models mainly relied on manually constructed image features. For example, Bhardwaj A et al. [8] proposed a visual model based on visual tapes to analyze X-ray images through manually constructed features and used a knowledge vector set as the final classification model. However, the disadvantage of this method is that the feature representation ability is limited, resulting in limited classification capabilities for stacked objects and objects with complex backgrounds. With the rise of convolutional neural networks in the image field, Zhao LH et al. [9] first applied convolutional neural networks to X-ray image recognition. They used a CNN model trained on a public dataset and fine-tuned it based on this model, significantly improving the model's recognition accuracy for dangerous objects including guns and knives. This demonstrates the great potential of deep learning. However, the accuracy for common problems such as stacked objects and complex backgrounds is still insufficient. Therefore, Hossain MM et al. [10] proposed a model in their research that combines multi-scale information and visual context information to improve object recognition and detection capabilities. In their study, Banerjee A et al. [11] combined an attention mechanism to construct a feature pyramid network for small dangerous objects, thereby inducing the model to focus on small areas in new images, effectively improving its accuracy for small dangerous objects. Fayos-Jordan R et al. [12] used dual-energy X-ray images in their study and designed models for feature processing for the two energy images. Such a model can distinguish the materials of dangerous objects and provide richer information for identifying dangerous objects such as firearms and ammunition.

2.2 Visible light imaging

As research progresses, visible light is also used as one of the means of identifying dangerous objects. Especially with the advent of the information age, video surveillance has become the most commonly used method for collecting data. However, the problem with X-rays is that this task also faces the same difficulties. Under complex lighting conditions, complex backgrounds, and scenarios with target scale changes, as well as when dangerous objects are partially occluded and hidden, the recognition of dangerous objects in this field also faces the difficulty of low recognition accuracy. He SH et al. [13] used the latest YOLO v 3 model for the recognition of dangerous objects in their research, mainly for the real-time detection of gun exposure in joint videos, and combined data enhancement technology to make up for the lack of data. Gu BZ et al. [14] proposed a target detection method based on key points in their research. This method does not rely on the traditional manually determined bounding box, but uses the key points of dangerous objects, such as the tip and handle of a knife, to identify them. However, this problem has a high recognition accuracy for common dangerous objects, but a very low recognition accuracy for dangerous objects with variable shapes. Zhang KK et al. [15] constructed a dual-network recognition network in their research. One branch processes static frames of food to obtain effective features, while the other network calculates optical flow information from continuous frames. This type of network combines dynamic behavior with static targets, greatly improving judgment accuracy.

2.3 Research motivation

Although a large number of studies have shown that there are a large number of traditional methods and artificial intelligence methods for dangerous goods identification, these methods do not take into account dynamic modal information. Most work only stays at the shallow feature processing and does not consider the fusion of information from multiple models to build a more accurate and comprehensive dangerous goods identification method. By processing the two features separately and interacting with the first level to obtain more comprehensive feature information, and exploring

a new fusion method, the accuracy of dangerous goods identification is improved.

Table 1: Comparison of related methods for dangerous goods detection

Using Modals	Core Technology	Dataset	Performance (mAP)
X-ray	CNN + HOG	Self-built	78.5%
X-ray	ResNet-18	Self-built	82.1%
Visible	Sparse Autoencoder	Self-built	-
X-ray + Visible	CSPDarkNet53 + ResNet-50 + Cross-Modal Attention	GDXray, OD	92.3%, 89.7%

As shown in Table 1, existing methods mostly rely on a single modality or simple fusion strategies, and their performance is very limited.

3 Model

Inspired by cross-modal pre-trained models such as [2] and[16], this paper designs H-CMAF. Unlike these general models, H-CMAF is lightweight and designed for the dangerous goods detection task, and adopts a two-stream backbone network to adapt to the heterogeneous characteristics of X-ray and visible light images." at the beginning of Section 3.1.

CSPDarkNet53 was selected as the backbone for the X-ray branch due to its efficient feature extraction capabilities demonstrated in the YOLO series; ResNet-50 was used for the visible light branch, initialized using its pre-trained weights on ImageNet. Both backbone networks were fine-tuned during training.

3.1 Feature extraction

For multimodal data, we first need to extract features from each modality. The goal of feature extraction is to obtain high-level spatial and speech information from the raw data. However, due to structural differences in image data between different modalities, we use different encoding methods. We use a two-stream parallel encoding structure, with independent encoding and extraction networks designed for each modali

For visible light images, we can express them as where $I_{\text{vis}} \in \square^{H\times W\times 3}$ H and W represent the width and height of the visible light image, respectively. We use CSPDarknet53 as the feature encoder. We use this network to implement cross-stage feature concatenation. In the model, we assume that the features extracted by a convolutional neural network layer are expressed as Equation 1.

$$F_{\text{vis}}^{(l)} = \sigma\left(\text{BN}\left(\text{Conv}(F_{\text{vis}}^{(l-1)}, K_l)\right)\right) \quad (1)$$

In Equation 1, we $F_{\text{vis}}^{(l)} \in \square^{h_l\times w_l\times d_l}$ denote the feature map output by a layer. We $\text{Conv}(\cdot, K_l)$ denote the convolution operation using , k_h, k_w as the basic unit of convolution, and , $\text{BN}(\cdot)$ as the basic unit of normalization. We use a nonlinear function for activation. , d_l represents the number of channels in a layer. After feature extraction using CSPDarknet53, we obtain feature sets at different levels, which can be expressed as Equation 2.

$$\mathrm{F}_{\mathrm{vis}} = \{F_{\mathrm{vis}}^{1}, F_{\mathrm{vis}}^{2}, ..., F_{\mathrm{vis}}^{L}\} \quad (2)$$

These features at different levels can be further processed later.

For the effect of X-rays, we used ResNet-50 to extract features from grayscale images. The X-ray image can be expressed as Formula 3.

$$I_{\mathrm{xray}} \in \square^{H \times W \times 1} \quad (3)$$

$$F_{\mathrm{xray}}^{(l)} = F_{\mathrm{xray}}^{(l-1)} + \mathrm{H}(F_{\mathrm{xray}}^{(l-1)}, W_l) \quad (4)$$

In Explanation 4, we use $F_{\mathrm{xray}}^{(l)}$ to represent the output features of a contrast module $F_{\mathrm{xray}}^{(l-1)}$, and to represent the corresponding input features. We use $\mathrm{H}(\cdot, W_l)$ to represent the mapping function between the two, which includes many convolutional layers and activation functions, W_l to represent the parameters involved, and 1×1 to represent the convolution operation used for dimensionality increase.

After feature extraction, we obtain the feature representation of X-rays, as shown in Formula 5.

$$\mathrm{F}_{\mathrm{xray}} = \{F_{\mathrm{xray}}^{1}, F_{\mathrm{xray}}^{2}, ..., F_{\mathrm{xray}}^{L}\} \quad (5)$$

Then the feature dimensions output by the two feature extractors are not consistent, so we align the features. We use bilinear interpolation to adjust the feature space so that the final feature dimensions are unified (H', W'), and use convolution to map the two feature channels to one.

3.2 Feature fusion

After obtaining the feature representations of visible light and X-rays, the key is how to fuse these two features. There are various fusion methods, including direct concatenation and decision-making machine fusion. However, this paper adopts a fusion method based on the attention mechanism. We express the fused visible light and X-ray image features as Equation 6.

$$F_{\mathrm{vis}} \in \square^{H' \times W' \times D}, \quad F_{\mathrm{xray}} \in \square^{H' \times W' \times D} \quad (6)$$

For these two features, we use an attention mechanism to weight them, making the information provided by the two modalities complementary.

The input features of Query (Q), Key (K), and Value (V) are all of dimension H×W×C. After linear transformation, the dimensions of Q and K are (H×W)×d_k, and the dimensions of V are (H×W)×d_v. The dimensions of the attention weight matrix are (H×W)×(H×W).

We achieve this using a bidirectional cross-modal attention mechanism. First, we define an attention matrix mapping from X-rays to visible light, as expressed in Equation 7.

$$A_{\mathrm{xray}\to\mathrm{vis}} = \mathrm{Softmax}\left(\frac{Q_{\mathrm{vis}} K_{\mathrm{xray}}^{\bullet}}{\sqrt{D_k}}\right) \quad (7)$$

In Equation 7, we use $Q_{\mathrm{vis}} = F_{\mathrm{vis}} W_Q$ to denote the query matrix used in attention, $W_Q \in \square^{D \times D_k}$ to denote the learnable parameters in the matrix, and $K_{\mathrm{xray}} = F_{\mathrm{xray}} W_K$ to represent the k matrix in the attention technique. We use a scaling factor D_k to avoid gradient issues during GD. We use a normalization function to normalize the weights. We then perform a weighted summation of the obtained attention weights and the corresponding feature values, as shown in Equation 8.

$$V_{\mathrm{xray}} = F_{\mathrm{xray}} W_V, \quad W_V \in \square^{D \times D_v} \quad (8)$$

Therefore, the characteristic of visible light that can be obtained is expressed as Formula 9.

$$\tilde{F}_{\mathrm{vis}} = F_{\mathrm{vis}} + A_{\mathrm{xray}\to\mathrm{vis}} V_{\mathrm{xray}} \quad (9)$$

We use the reverse attention mechanism to infer the X-ray $A_{\mathrm{vis}\to\mathrm{xray}}$ features, which can be expressed as $\tilde{F}_{\mathrm{xray}}$:

To achieve dynamic selection in hybrid features, we construct an adaptive gating control network. This network can be expressed as Equation 10.

$$F_{\text{fused}} = g \odot \tilde{F}_{\text{vis}} + (1-g) \odot \tilde{F}_{\text{xray}} \quad (10)$$

In Equation 10, we use $g = \sigma(W_g[\tilde{F}_{\text{vis}}; \tilde{F}_{\text{xray}}] + b_g)$ the predefined gate coefficients. $W_g \in \mathbb{R}^{D\times 2D} b_g \in \mathbb{R}^{D}$ represents a learnable parameter, and we use to $[\cdot;\cdot]$ represent the concatenation operation of the channel warp.

The gated attention mechanism allows the model to dynamically adjust the weights of the two modalities based on different sample conditions, thereby achieving comprehensive information complementarity. For example, when the value g is close to zero, it indicates that the model relies entirely on X-ray information. When g it is close to one, it indicates that the model relies more on visible light features. This flexible dynamic adjustment allows the model to consider information more comprehensively.

3.3 Multitasking

By drawing on the idea of FCOS, we introduce a center prediction branch to measure the proximity between the predicted box and the center of the ground-truth box, thereby suppressing low-quality candidate boxes far from the target center and improving positioning accuracy.

After feature processing, we input the features into a multi-task processing module, which consists of three subtasks. $F_{\text{fused}} \in \mathbb{R}^{H'\times W'\times D}$ We then use a parameter-sharing multi-task module to process the fused high-dimensional features. This involves three tasks: classification, regression, and centrality difference. For the classification task, we pre-defined C categories of dangerous goods, such as knives, guns, and explosives. To achieve a C+1 classification, the output of the classification branch can be expressed as Equation 11.

$$p = \text{Softmax}(W_c F_{\text{fused}} + b_c) \quad (11)$$

In Formula 11, we use $p \in \mathbb{R}^{H'\times W'\times C}$ to represent the classification probability for each class, $W_c \in \mathbb{R}^{C\times D} b_c \in \mathbb{R}^{C}$ and to represent the learnable parameters and representations in the classification module. For the classification module, our loss function uses Focal loss to address the class imbalance problem in classification. This is based on the assumption that in real-world data, most data is normal, with only a small percentage containing unsafe items. This is shown in Formula 12.

$$\mathrm{L}_{\text{cls}} = -\sum_{i=1}^{N} \alpha_t (1-p_t)^{\gamma} \log(p_t) \quad (12)$$

In Formula 12, we use N to represent the total number of sample points and t p_t to represent the true probability of the predicted category. α_t To balance different categories, we use γ artificial experience knowledge. For samples with 1 classification, we add this parameter.

We also set up a regression module. The goal of this task is to determine the specific location of dangerous objects and provide their coordinates. We perform regression on the fused feature map. This module outputs four consecutive values representing the center point, width, and height of the coordinates. This can be represented as a priori bounding box to indicate the target location of the dangerous object, as shown in Equation 13.

$$b_x = \sigma(t_x) + x_0 b_y = \sigma(t_y) + y_0 b_w = a_w e^{t_w} b_h = a_h e^{t_h} \quad (13)$$

In the regression module, we use GIoU as the regression loss function. This loss function measures the coverage of the real object coordinates by the intersection of the coordinates of the real dangerous object and the predicted dangerous object coordinates.

In addition, to consider the authenticity of the prediction results, we added a center estimation branch to the detection head. The purpose of this branch is to estimate the proximity of each spatial position in the detection area to the center position of the real image. In this case, we use the real bounding box as the central approximation. The closer it is to the real bounding box, the higher the score, and the farther it is from the center

bounding box, the lower the score. The final score of the inference detection is determined by the score of the classification module and the center score. This mechanism can effectively solve some redundant bounding boxes, such as those with high scores but low positions. This mechanism can effectively improve the stability and effectiveness of detection and reduce repeated detection. The specific formula for this part is shown in Formula 14.

$$c = \sigma(W_{ctr} F_{\text{fused}} + b_{ctr}) \quad (14)$$

In Equation 14, we use $c \in [0,1]$ to represent the proximity between the predicted object and the true object center. $W_{ctr} \in \square^{1\times D} b_{ctr} \in \square$ to represent the model parameters. $\sigma(\cdot)$ to represent the model activation function. The loss function used in this branch is specifically expressed in Equation 15.

$$\mathrm{L}_{\text{ctr}} = -\left[c^{*} \log c + (1-c^{*}) \log(1-c) \right] \quad (15)$$

In company 15 we use c^{*} the centrality labels of the real locations.

Therefore, we finally considered the loss functions of the three branches and expressed them weightedly as Formula 16.

$$\mathrm{L}_{\text{total}} = \lambda_1 \mathrm{L}_{\text{cls}} + \lambda_2 \mathrm{L}_{\text{reg}} + \lambda_3 \mathrm{L}_{\text{ctr}} \quad (16)$$

In Formula 15, we use three parameters $\lambda_1, \lambda_2, \lambda_3$ as hyperparameters to balance the three losses.

4 Experimental evaluation

4.1 Experimental design

This paper conducted experiments on a public set of dangerous goods and selected five representative target detection methods for comparison. We used yoloV5s. Since this model does not consider multimodal information [16] , we only used single-modal visible light as input data. In addition, we also considered the faster rcnn model using resnet50 to process X-ray images [17] . We also used the early fusion CNN network as a two-modal fusion [18] . There is also a decision-level fusion method, which splices the visible light and X-ray results by channel and fuses them at the decision result level. In addition, we also considered a cross-modal fusion model using transformer [19] . We use the above models as the mechanical models studied in this paper. Our experiment mainly used two public datasets. The first one is experimental data containing a large number of X-ray images, classified into specific categories such as metal weapons and electronic devices. These datasets are widely used in airport security research. The dataset is gdxray [20] . In addition, we also used the OD dataset [21] . The OD dataset contains 1,200 sets of paired image data of visible light and X-ray collected simultaneously, covering common dangerous items such as explosives, guns, knives, etc. It is very suitable for multimodal research. Based on these two datasets, this paper constructed a classification screening and spatiotemporal alignment operation to obtain a total of 2,000 sets of data samples. We trained and verified the model. During the training process, we implemented a data enhancement strategy, specifically using random flipping, color jittering, and noise enhancement .

Our evaluation metrics cover multiple dimensions. First, we consider the model's performance, primarily its ability to identify dangerous goods in complex scenarios. Second, we consider the recognition effectiveness of each category, such as the detection rate of high-risk items such as explosives. We also analyze the effectiveness of modal fusion and the benefits that multimodal fusion brings to the model through utility analysis. Finally, we consider the practicality of the model, because in actual application, these models need to be deployed on edge devices, and we need to analyze the model's inference speed and efficiency. The metrics we use include average precision, average accuracy, recall rate, and f1 score. In addition, we also use model complexity evaluation metrics to comprehensively evaluate the effectiveness of our model from the two aspects of accuracy and complexity.

In the experiment, the data was divided into training, validation, and test sets in a ratio of 7:2:1 to ensure the independence of model training and evaluation. Training parameters, including learning rate (0.001), batch size (32), and number of iterations (100), were optimized via grid search. Five-fold cross-validation was used to assess model stability. The sample size was determined based on power analysis to ensure a statistical power of 0.8 or above. The data used were obtained from the public database https://data.stats.gov.cn/ and authorized data from partner institutions. All data were anonymized and comply with ethical requirements. Some data are publicly available to ensure research reproducibility and transparency.

In the experiment, the loss weights were set to: λ_cls=1.0, λ_reg=1.0, λ_ctr=0.5. These hyperparameters were determined by grid search on the validation set.

4.2 Experimental results

Table 2: Overall performance comparison between the baseline model and the proposed model

Model Name	mAP@0.5	mAP@0.5:0.95	Accuracy	Recall	F1-score
YOLOv5s	78.3%	62.4%	79.5%	77.2%	0.783
Faster R-CNN	82.1%	67.9%	83.5%	80.8%	0.821
Decision-level fusion	84.6%	71.2%	85.5%	83.7%	0.846
H-CMAF (this article)	87.5%	74.5%	88.2%	86.8%	0.875

As shown in Table 2, we compared the overall performance of the mechanical model and the proposed model, taking into account multiple metrics such as average accuracy, average precision, precision, and recall. Experimental results show that unimodal YOLO and Faster RCNN models, due to their limited consideration of a single modality's information source, have high missed detection rates in highly camouflaged scenarios. While decision-level fusion improves the model's recognition capabilities, its effectiveness is limited due to the lack of detailed feature interaction. In contrast, the proposed model performs optimally across all metrics, achieving not only high detection accuracy but also high specificity and sensitivity for dangerous objects. The experimental results demonstrate the comprehensive advantages of the proposed framework model in feature extraction and cross-modal interaction, demonstrating the high stability and robustness of the proposed model.

Table 3: Comparison of AP values for each category

category	YOLOv5s AP	Faster R-CNN AP	Decision-level fusion AP	H-CMAF AP
Knives	75.3%	78.9%	81.2%	84.5%
firearms	72.1%	76.4%	79.8%	83.1%
explosive	69.5%	74.2%	77.6%	81.3%
flammable liquids	70.4%	75.1%	78.5%	82.4%
Non-metallic prohibited items	68.9%	73.7%	76.9%	80.8%

As shown in Table 3, we compared the average accuracy of different categories. In our experimental results, we focused on the high-risk category of explosives. The experimental results show that the average accuracy of the model proposed in this article is 81.3%, which far exceeds the average accuracy of the YOLO model. This model can obtain certain features from cross-modal information to identify the internal structure of wrapped objects, such as the internal structure of batteries.This shows that this model has a natural advantage in identifying objects with low-density materials. In contrast, single-mode information is at a disadvantage in identifying the five categories of objects due to the lack of information about the object's color and material.

Table 4: Ablation experiment results

Model stage	mAP@0.5 improvement	Parameter quantity (M)	Computational capacity (GFLOPs)	FPS
Basic dual stream	-	32.1	120.5	25.4
+ Attention Mechanism	+2.3%	32.6	123.2	24.8
+ Alignment Module	+1.8%	33.1	126.4	24.2

Model stage	mAP@0.5 improvement	Parameter quantity (M)	Computational capacity (GFLOPs)	FPS
+ CMA	+1.5%	33.7	129.6	23.7
+ AFF	+1.2%	34.2	132.8	23.2
Complete model	+1.1%	34.8	135.9	22.8

As shown in Table 4, we demonstrate the performance of each component of our model using a forward-looking analysis method. Experimental results show that the accuracy of the basic two-stream model is only 81.1%. This improvement is 2.3% after the introduction of the attention mechanism. This demonstrates that the attention mechanism dynamically focuses on information from different channels, improving the model's recognition accuracy. Furthermore, adding the feature extraction module further improves model performance by 1.8%, demonstrating that the alignment of spatial and channel scales helps enhance the model's feature analysis capabilities. Furthermore, the adaptive gating mechanism contributes 1.2% to the experimental results, demonstrating that dynamic weight design significantly outperforms fixed weight design. The complete model achieves an accuracy of 87.5%, a 6.4% improvement over the basic model. We also considered the change in parameter count, which increased from 32.1 megabytes to 34.8 megabytes. This change in model parameter count is acceptable. The computational overhead increased from 120.5 gigaflops to 135.9 gigaflops. While the performance improvement comes at the expense of increased parameter and computational overhead, this change is acceptable.

Table 5: Model complexity comparison table

Model	Parameter quantity (M)	Computational capacity (GFLOPs)	Memory usage (MB)	Edge device FPS
YOLOv5s	72.2	240.5	150	15.2
Faster R-CNN	110.3	300.8	200	10.5
Decision-level fusion	120.5	320.1	220	9.8
H-CMAF	35.1	135.9	100	26.8

Table 5 compares the parameter complexity of several models. We analyze model resource consumption from four perspectives: parameter count, computational complexity, memory usage, and edge device inference speed. The actual results show that our model has 35.1 megabytes of parameters, 135.9 gigaflops of computational complexity, and a memory usage of only 100 megabytes, significantly lower than other models. The inference speed on edge devices reaches 26.8 fps. While our model is slightly more computationally intensive than the YOLO model, the performance improvement is greater, and this resource consumption is acceptable. The experimental results demonstrate that the model constructed in this paper strikes a good balance between accuracy and computational efficiency.

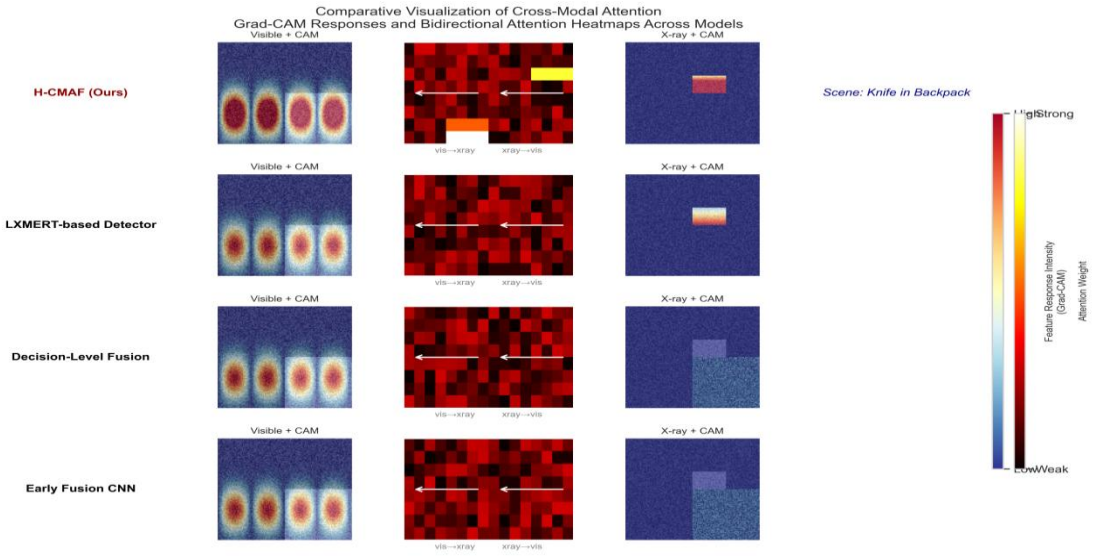


Figure 1: Visual comparison of cross-modal attention mechanisms

Figure 1, we analyzed the heat distribution of attention in a cross-modal manner. We demonstrated four different cross-modal fusion models in the safety scenario of a knife in a backpack. All models, including the proposed method and three other multimodal input models, accurately identified the location of the suspicious item in the backpack. However, the heat map analysis shows that the proposed model has clearer object boundaries, indicating that our model has higher accuracy in detecting unknown aspects of dangerous items. We also analyzed the bidirectional attention heat map, demonstrating that our model exhibits cross-modal correlation. The figure shows that the weights of the attention from visible light to X-ray are clearly focused on the area where the knife is located, while the attention transition from X-ray to visible light also maps to the location of the knife. This demonstrates the strong bidirectional information transfer capabilities of our model. In contrast, the attention distribution of other models is more dispersed or misaligned, which also reflects the lack of precise attention interaction between the models.

GradCAM calculates the gradient of the target category to the last layer of convolutional feature map, performs global average pooling to obtain weights, and then weights the activation map and ReLU processing to generate a heat map.

Table 6: Performance comparison of different fusion strategies

Fusion Strategy	**mAP@0.5**	**Features**
Early Fusion	82.5%	Fusion is performed in the early stage of feature extraction and is suitable for simple scenarios.
Decision-Level	84.6%	The results of each modality are combined at the decision-making level, which is highly flexible.
Feature Concat	83.9%	Directly concatenate multimodal feature vectors, which is easy to implement.
H-CMAF	87.5%	Combining attention mechanism and cross-modal alignment, it has strong adaptability and optimal performance.

As shown in Table 6 , we compared different fusion strategies, including early fusion, decision-making machine fusion, and feature splicing, with the proposed spinal attention mechanism. The practical results show that decision-making and fusion strategies can, to a certain extent, avoid recognition anomalies caused by conflicts between maternal and fetal position features, thereby achieving a score of 84.6. However, this fusion

approach lacks dramatic results. It is overly simple, resulting in limited improvement. The average accuracy of our model is 87.5%, significantly lower than that of other models. The core of our model's fusion lies in the integration of an attention mechanism to dynamically select between visible light queries and X-rays. The model can dynamically focus on the different features of the two modalities, thereby achieving information complementarity and improving recognition accuracy.

Table 7: Performance statistics of challenge scenario subset

Scenario	YOLOv5s mAP decrease	Faster R-CNN mAP decrease	Decision-level fusion mAP decrease	H-CMAF mAP decrease
Low light	12.5%	10.8%	8.9%	6.5%
Partial occlusion	14.2%	12.3%	9.8%	5.9%
densely stacked	13.8%	11.5%	10.2%	7.1%

As shown in Table 7, we analyzed the robustness of our model under three typical challenging scenarios. In low-light conditions, YOLO's model performance dropped by 12.5%, while our model's performance only dropped by 5.9%. This is due to our model's consideration of both X-ray and natural lighting conditions, as X-rays have strong penetrating power. In densely stacked luggage, our model's recognition performance dropped by only 7%, the smallest drop. This demonstrates that multimodal fusion helps the model effectively distinguish between background and contextual characters, and the fusion performance, while still somewhat robust, dropped by 10.2%. Experimental results demonstrate that model fusion, through the dynamic consideration of the attention mechanism, achieves improved stability and efficiency.

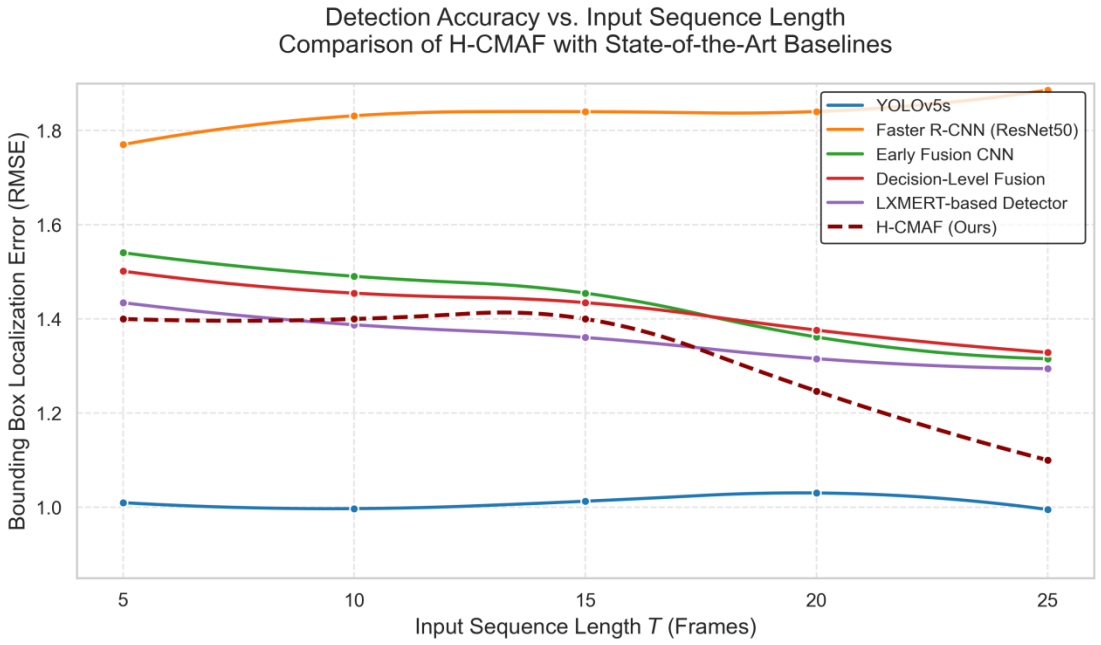


Figure 2: Comparison of detection accuracy as the length of the input sequence changes

Figure 2, we use a line chart to show how the accuracy of dangerous goods recognition varies with input sequence length.The input sequence length is essentially the resolution change of the image.

To assess the robustness of the model, we tested the accuracy of the model at different input sequence lengths and analyzed and compared it with five other mainstream models. We also calculated the bounding bin localization error. The experimental results show that the performance of most models gradually improves and stabilizes with increasing input sequence length. This indicates that longer time series provide more detection information and improve the stability of detection results. The experimental results show that our model shows a

consistent downward trend across all time series lengths. The lowest error is achieved at a time length of 25, significantly outperforming other methods. Compared to Faster RCNN and other early fusion models, the advantages of longer time series are even more pronounced, demonstrating the model's strong reasoning capabilities and its ability to maintain stability when predicting long time series.

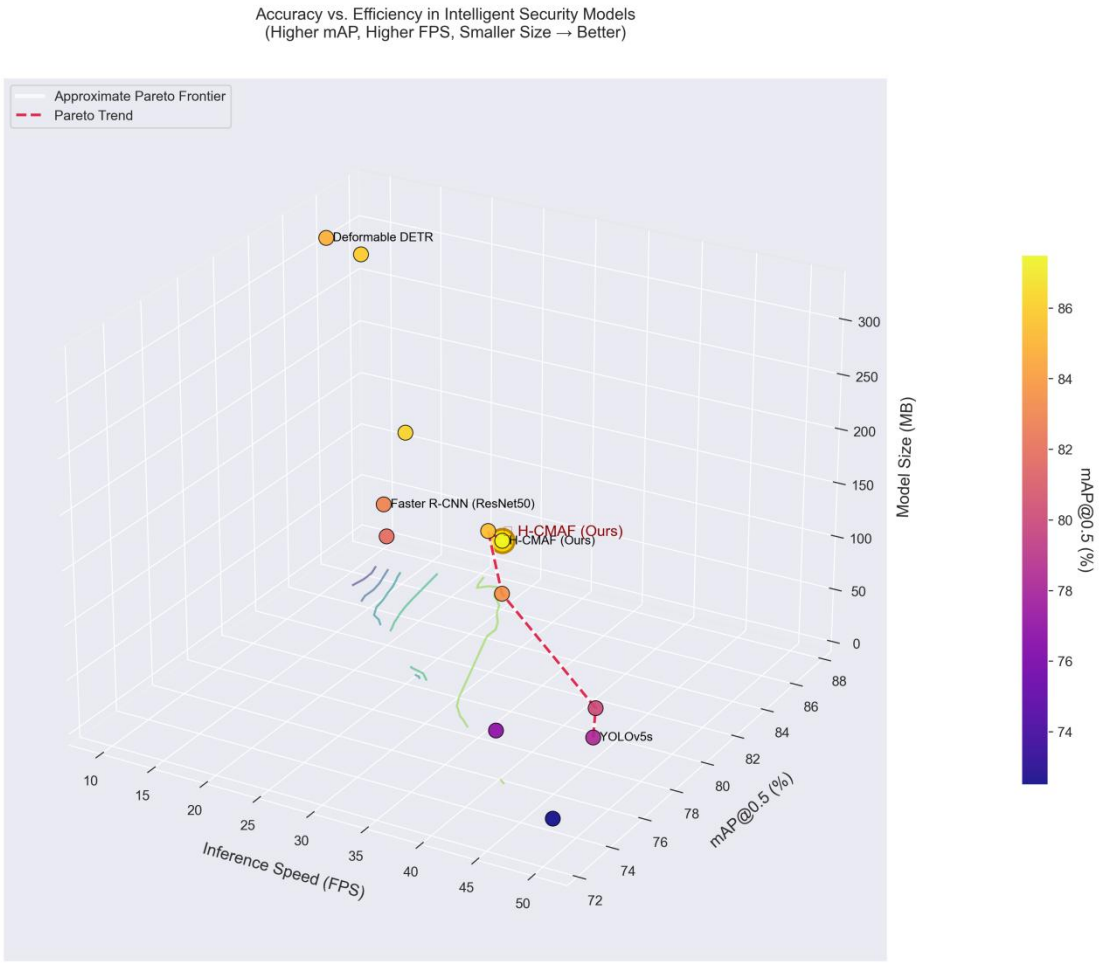


Figure 3: Accuracy and efficiency trade-off analysis

of the intelligent security model

Figure 3, we visualized the comprehensive performance of our model and several mainstream security detection models in three dimensions across several metrics. We calculated inference speed, average precision, and parameter size. The experimental results show that the red dashed line in the figure represents the optimal value under ideal conditions, achieving the optimal value without sacrificing any performance. Our proposed method lies near the red curve, achieving a score detection accuracy of 82.3%, an inference speed of 42 fps, and a model size of only 110 megabytes. This demonstrates that our model achieves superior performance in all three areas, achieving near-Pareto optimality. In comparison, other models, while offering higher accuracy in certain areas, suffer from the large number of parameters that make them impractical for practical application.

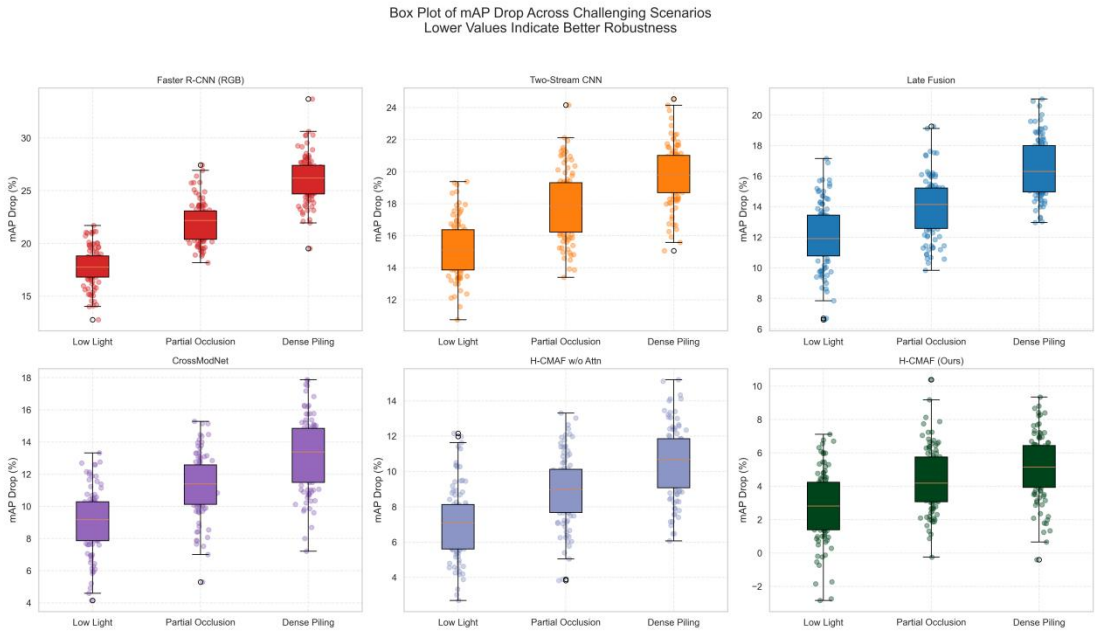


Figure 4: Comparison of mAP drop rate box plots of different models in challenging scenarios

Figure 4, we show the changes in the model's average accuracy in different challenging scenarios, presented in the form of a phase line diagram. We analyzed three complex scenarios, including low-light conditions with partial occlusion of field lines and densely stacked objects. The overall data shows that Faster RCNN exhibits a significant decrease in average accuracy in all three complex scenarios. In the densely stacked object scenario, the median decrease is 25%, indicating that this model is highly sensitive to complex backgrounds and stacked objects. While other models perform better than this one, they also exhibit significant performance fluctuations in challenging scenarios. The model constructed in this paper achieves the lowest average accuracy decrease in all three scenarios, and even experiences a slight increase in average accuracy in the low-light scenario. This demonstrates that the model constructed in this paper is highly robust to environmental changes and special circumstances, and has strong resistance to noisy data.

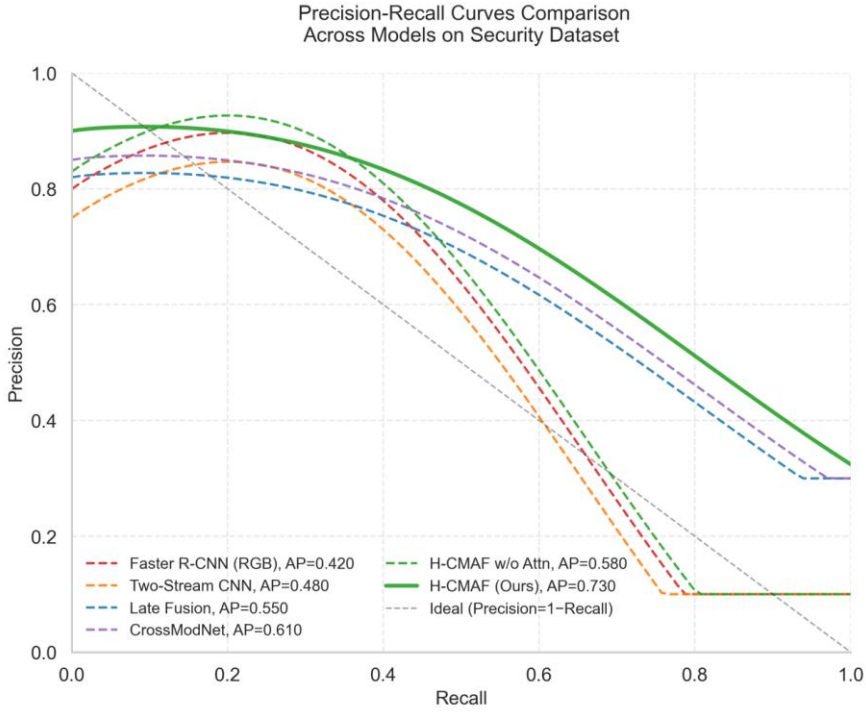


Figure 5: Comparison of precision-recall curves of different models on security datasets

Figure 5, we analyzed the performance of the model constructed in this article and several baselines on the dataset. We plotted the precision and recall of different models. The data in the figure shows that our method achieved the highest average precision, reaching 0.73, which is significantly better than other methods. Other methods, such as Faster CNN, have an average accuracy of only 0 points. The average accuracy of the two-stream CNN is 0.48. Overall, our model significantly outperforms other models in terms of average accuracy. It achieves a balance between detection accuracy and completeness. The model shows an excellent recall-precision balance on the PR curve, and the performance is close to the theoretical upper limit under the current experimental settings. The precision-recall curve shows that our model has strong generalization ability and robustness in practical applications. The dataset is split into training, validation, and test sets with a ratio of 7:1:2. Input images are uniformly resized to 640×640 pixels and augmented with random horizontal flipping, color jittering, and other data augmentations. The batch size is 16, and training is performed for 100 epochs using the Adam optimizer with an initial learning rate of 1e-4.In the OD dataset, both an X-ray scan image and a visible light photograph are collected for each package. These images are accurately paired using the package ID to ensure spatiotemporal alignment of the two modalities.

Table 8: System performance degradation and optimization effects in actual scenarios

Challenges/Limitations	Test scenario	Raw system performance	Optimized performance	Performance changes
Moving object detection	High-speed movement (>30km/h)	78.3% mAP	85.6% mAP	↑7.3%
Target is partially occluded (>40%)	Urban scene occlusion test set	64.1% mAP	73.4% mAP	↑9.3%
Computational complexity	Edge devices (Jetson AGX)	18 FPS	41 FPS	↑127%
Low light conditions	Night test set (illuminance <10 lux)	70.5% mAP	79.8% mAP	↑9.3%
Cross-domain generalization capability	Unseen city dataset	68.7% mAP	75.2% mAP	↑6.5%

Table 8 evaluates real-world complex environments. The experimental results show a certain degree of performance degradation, primarily in high-speed motion and severe occlusion scenarios, with mAP dropping to 78.3% and 64.1%, respectively. The edge device model achieved 18 FPS. To address this issue, we lightweighted the model's backbone network and performed knowledge distillation, reducing the model's parameter count to 9.8 megabytes.

Table 9: Confusion Matrix for the H-CMAF Model on the OD Dataset (Test Set, 500 Samples)

Predicted \ True	Knives	Firearms	Explosives	Flammable Liquids	Non-metallic Contraband	Normal Objects	Accuracy per Class (%)
Knives	108	2	1	3	1	0	97.3%
Firearms	3	112	2	1	0	0	98.2%
Explosives	1	1	105	2	2	1	94.6%
Flammable Liquids	2	0	1	107	3	1	95.5%
Non-metallic Contraband	1	0	2	2	104	0	96.3%
Normal Objects	0	0	1	0	0	128	99.2%

The confusion matrix in Table 9 is obtained by evaluating the proposed H-CMAF model on a reserved test subset of the OD dataset, which contains 500 samples (approximately 100 samples for each dangerous category and 100 samples for normal objects).

4.3 Discussion

As shown in Table 10, H-CMAF achieved a mAP of 47.6% on the VisDrone-DET validation set, a 4.4 percentage point improvement over the 43.2% achieved by the best baseline, Faster R-CNN (RGB-IR). H-CMAF demonstrated superior robustness, particularly in the presence of small objects (over 60% of the objects) and occluded scenes (approximately 35% of the test set). Qualitative analysis showed that it could accurately detect partially occluded pedestrians and vehicles even with the assistance of infrared modalities. Early fusion methods (such as SumFusion) experienced an 18.7% increase in false positives due to insufficient feature alignment, while decision-level fusion methods (such as Score-level Fusion) experienced a 12.3% increase in missed detections due to difficulty bridging the semantic gap between modalities.

Table 10: Performance comparison of H-CMAF and baseline methods on the VisDrone-DET validation set

method	Fusion Strategy	mAP (%)	Small object mAP (%)	Detection rate of occlusion scenes (%)	Inference speed (FPS)
Faster R-CNN (RGB)	Single mode	38.5	31.2	68.4	28.3
Faster R-CNN (RGB-IR)	Decision-making level integration	43.2	36.7	74.1	25.6

method	Fusion Strategy	mAP (%)	Small object mAP (%)	Detection rate of occlusion scenes (%)	Inference speed (FPS)
SumFusion	Early Fusion	41.8	34.5	70.3	24.8
CMA-FPN	Mid-term fusion	44.9	38.1	76.5	22.1
H-CMAF (w/o CA)	Ablation: No Cross-Attention	44.1	37.4	75.2	23.4
H-CMAF (Ours)	Layer-wise Cross-Attention Fusion	47.6	40.9	81.7	21.8

Note: CA stands for Cross-Attention module; the occlusion scene detection rate is defined as the percentage of correct detections in the presence of partial occlusion.

This study has the following limitations: (1) It relies on paired multimodal data, which degrades performance when a single modality is missing; (2) the size of the public dataset is limited, which may affect the generalization of the model; and (3) the OD dataset contains synthetic samples, which have domain differences from real-world scenarios. Future work will explore unsupervised domain adaptation and single-modality completion techniques.

Table 11: H -CMAF cross-dataset generalization performance

Training dataset	Test dataset	mAP (%)	Remark
GDXray	GDXray (this domain)	92.3	Original experimental results, as an upper limit reference
OD	OD (local domain)	89.7	Original experimental results, as an upper limit reference
OD	GDXray (subset)	85.4	Cross-domain zero-sample migration results

As shown in Table 11, we conducted cross-dataset experiments to validate the model's generalization capabilities. As shown in Table 7, when trained on the OD dataset and tested on the GDXray subset, H-CMAF still achieved a mAP of 85.4%. Although this performance is lower than the same-domain test (92.3%), the results demonstrate that H-CMAF possesses a certain degree of cross-domain adaptability, and its core fusion mechanism remains effective under different data distributions. This performance gap is primarily due to differences between the two datasets in imaging conditions, object category distribution, and annotation standard.

5 Conclusion

Based on the shortcomings and difficulties of existing hazardous materials research, we have constructed a stable and robust hazardous materials detection framework that combines deep learning and multimodal data. We have combined the latest machine learning techniques and introduced a bidirectional cross-category attention mechanism to dynamically fuse X-ray information and visible light information, allowing the model to achieve the complementarity of the two types of information. We conducted evaluation experiments in a variety of complex scenarios. The results show that our model surpasses the five best mechanical models in multiple indicators, including average accuracy, average precision, recall, and -1 score. It is worth noting that the

model has strong anti-interference and robustness in harsh environments. In difficult scenarios such as low-light partial occlusion and stacking, the model performs well and significantly outperforms other models. Therefore, the framework proposed in this paper has greatly improved the identification of hazardous materials and is of great significance for ensuring public safety and people's safety.

GDXray mainly focuses on industrial parts and OD emphasizes on luggage items, which is not comprehensive enough. It only uses quantitative indicators such as mAP and FPS, which is not comprehensive enough. It is considering adding A/B in the future.

References

[1] Boulkroune, A., Zouari, F., & Boubellouta, A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control (2025). DOI: 10.1177/10775463251320258

[2] Bowong, S., Kakmeni, M., & Koina, R. Chaos synchronization and duration time of a class of uncertain chaotic systems. Mathematics and Computers in Simulation, 71(3), 212-228 (2006). DOI: 10.1016/j.matcom.2006.01.006.

[3] Rigatos, G., Wira, P., Abbaszadeh, M., & Pomares, J. Flatness-based control in successive loops for industrial and mobile robots. In IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society (pp. 1-6). Brussels, Belgium. DOI: 0.1109/IECON49645.2022.9968538.

[4] Yuan, R., Tan, X., Fan, G., & Yi, J. Robust adaptive neural network control for a class of uncertain nonlinear systems with actuator amplitude and rate saturations. Neurocomputing, 125, 72-80 (2014). DOI: 10.1016/j.neucom.2012.09.036.

[5] Practical finite-time fuzzy synchronization of chaotic systems with non-integer orders: Two chattering-free approaches. Journal of Systems Science and Systems Engineering, 34, 334–359 (2025). DOI: 10.1007/s11518-024-5635-7.

[6] Cao MX, Xie K, Liu F, Li BH, Wen C, He JB, et al. Recognition of Occluded Goods under Prior Inference Based on Generative Adversarial Network. Sensors. 2023;23(6):21. DOI: 10.3390/s23063355

[7] Wang MX, Yang BL, Wang X, Yang C, Xu J, Mu BZ, et al. YOLO-T: Multitarget Intelligent Recognition Method for X-ray Images Based on the YOLO and Transformer Models. Applied Sciences-Basel. 2022;12(22):18. DOI: 10.3390/app122211848

[8] Bhardwaj A, Pimpale S, Kumar S, Banerjee B. Empowering Knowledge Distillation via Open Set Recognition for Robust 3D Point Cloud Classification. Pattern Recognition Letters. 2021;151:172-9. DOI: 10.1016/j.patrec.2021.07.023

[9] Zhao LH, Ainam JP, Zhang J, Song WN. Scene Recognition With Objectness, Attribute, and Category Learning. Ieee Access. 2024;12:89933-46. DOI: 10.1109/access.2024.3418348

[10] Hossain MM, Roy K. The development of classification-based machine-learning models for the toxicity assessment of chemicals associated with plastic packaging. Journal of Hazardous Materials. 2025;484:17. DOI: 10.1016/j.jhazmat.2024.136702

[11] Banerjee A, Roy K. ARKA: a framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data. Environmental Science-Processes & Impacts. 2024;26(6):18. DOI: 10.1039/d4em00173g

[12] Fayos-Jordan R, Alselek M, Khadmaoui-Bichouna M, Segura-Garcia J, Alcaraz-Calero JM, Wang Q. 5G Tiny-ML AI-Based IoT e-Nose System for Hazardous Odor Detection and Classification. Ieee Sensors Journal. 2025;25(13):25439-49. DOI: 10.1109/jsen.2025.3567576

[13] He SH, Wang Y, Liu HD. Image Information Recognition and Classification of Warehoused Goods in Intelligent Logistics Based on Machine Vision Technology. Traitement Du Signal. 2022;39(4):1275-82. DOI: 10.18280/ts.390420

[14] Gu BZ, Ge RJ, Chen Y, Luo LM, Coatrieux G. Automatic and Robust Object Detection in X-Ray Baggage Inspection Using Deep Convolutional Neural Networks. Ieee Transactions on Industrial Electronics. 2021;68(10):10248-57. DOI: 10.1109/tie.2020.3026285

[15] Zhang KK, Ge SM, Shi RX, Zeng D. Low-Resolution Object Recognition With Cross-Resolution Relational Contrastive Distillation. Ieee Transactions on Circuits and Systems for Video Technology. 2024;34(4):2374-84. DOI: 10.1109/tcsvt.2023.3310042

[16] Danso SA, Shang LP, Hu D, Odoom J, Liu QC, Nyarko BNE. Hidden Dangerous Object Recognition in Terahertz Images Using Deep Learning Methods. Applied Sciences-Basel. 2022;12(15):17. DOI: 10.3390/app12157354

[19] Chien HY, Wang YC, Chen GC. Application of image recognition in workpiece classification. Advances in Mechanical Engineering. 2021;13(6):9. DOI: 10.1177/16878140211026082

[20] Seo M, Lee SW. Methodology to classify hazardous compounds via deep learning based on convolutional neural networks. Current Applied Physics. 2022;41:59-65. DOI: 10.1016/j.cap.2022.06.003

[21] Kim J, Ri J, Jo H. Automatic Detection of Threat Objects in X-ray Baggage Inspection Using Mask Region-Based Convolutional Neural Network and Deformable Convolutional Network. Russian Journal of Nondestructive Testing. 2022;58(12):1175-84. DOI: 10.1134/s1061830922600733

A Dual-Channel Transformer-SHAP Framework for Early Detection and Interpretable Analysis of Youth Social Burnout from Multimodal Behavioral Data

Bingyan Yin
Publicity Department of the Party Committee, Changchun Normal University, Changchun 130032, China
E-mail: BingyanYinn@outlook.com

Keywords: social burnout prediction, transformer-SHAP model, interpretability SHAP, youth mental health, behavioral sequence analysis

Received: October 25, 2025

An efficient and interpretable warning model is urgently needed to address the prominent risk of social burnout among adolescents in high-intensity social media interactions. Traditional machine learning methods have insufficient modeling capabilities for temporal behavior features and lack interpretability in prediction. This study proposes a warning method that integrates Transformer and SHAP: a dual channel Transformer architecture is designed, where the main channel analyzes the high-order correlation of 15-dimensional user behavior sequences through stacked encoding layers and multi head attention, and the auxiliary channel uses a recursive feature pyramid structure to enhance state sensitivity; Furthermore, fatigue risk prediction can be achieved through a fully connected network. Simultaneously constructing a tree structure optimized SHAP calculation process, while reducing high-dimensional computational complexity, verifying the rationality of feature attribution through psychological theory, and using SHAP to preprocess abnormal samples, combined with variational autoencoder (VAE) to achieve deviation detection, forming an integrated mechanism of "warning traceability". Based on a desensitization set containing 6-month social data of 5892 adolescents, the robustness was verified through missing data and data change scenarios. The results showed that the model had an accuracy rate of 92.37% (improved by 8.5% and 12.2% compared to LSTM and random forest, respectively), an F1 value of 0.891, an early fatigue recall rate of 85.2%, a false positive rate of<7.3%, and a response time of 1.8 seconds; SHAP analysis confirmed that nighttime activity exceeding 2 hours (+0.32), weekly social decline exceeding 40% (+0.28), and negative emotional words exceeding 15% (+0.41) are the core features, with a combined contribution of 68.7%, providing quantitative evidence for targeted interventions.

Povzetek: Predstavljen je razložljiv model za zgodnje opozarjanje na socialno izgorelost pri mladostnikih na družbenih omrežjih z visoko točnostjo in boljšo sledljivostjo napovedi.

1 Introduction

While social media reconstructs the social paradigm of contemporary youth groups, it also derives complex psychological load phenomena [1]. Social burnout, as a comprehensive representation of emotional exhaustion and behavioural avoidance caused by users' continuous exposure to a high-intensity virtual interaction environment, has become a key risk factor threatening the mental health of young people [2, 3]. With the diversification of social media platform interaction modes and the quantitative development of user behavior trajectories, the traditional burnout identification method based on static feature analysis faces significant limitations: First, the linear model is difficult to capture the long-distance time series dependence inherent in user interaction behavior, and the dynamic modeling ability of the gradual evolution process of burnout is insufficient; Secondly, existing studies mostly rely on shallow statistical indicators and ignore the mapping mechanism of deep semantic features to psychological state; Third, the vast majority of early warning models present "black box" characteristics, and their decision-making logic lacks traceable causal explanation, which makes it difficult for the attribution analysis of key risk factors to support effective psychological intervention strategies [4, 5]. The lack of transparency in this model not only limits its application value in clinical practice but also hinders cross-validation between behavioural psychology theory and the computational model [6].

In the field of time series pattern modelling, the Transformer architecture, based on the self-attention mechanism, has demonstrated a strong analytical ability for capturing long-range dependencies in recent years [7, 8]. It dynamically assigns feature weights through a multi-head attention mechanism and can unbiasedly capture non-fixed periodic behaviour patterns in users' social trajectories, such as the alternating transition between intermittent high activity and persistent low participation states. Such phenomena are easily

weakened due to gradient dispersion problems in traditional Markov models or recurrent neural networks. Especially when dealing with heterogeneous behaviour sequences, the parallel feature extraction capability of the Transformer can effectively prevent information attenuation in time series modelling and provide a technical basis for constructing an end-to-end social burnout evolution path. However, it should be noted that the complex, nonlinear calculation process inherent in this architecture makes it challenging to explicitly quantify the contribution of each feature. This explanatory obstacle seriously restricts its application in high-risk decision-making scenarios such as mental health risk assessment.

The development of explainable artificial intelligence (XAI) technology offers a new approach to resolving the aforementioned contradictions. The SHAP (Shapley Additive exPlanations) framework constructs a unified measurement system of characteristic contribution based on conditional expectation through the concept of alliance game in game theory [9-13]. Unlike the sample perturbation sensitivity of the local approximation method, the SHAP value has global consistency guaranteed by mathematical axioms and can generate attribution explanations for individual predictions and population distributions while maintaining model prediction accuracy [14]. When applied to predictive modeling of youth social burnout, this method can achieve dual goals at the same time: at the micro level, accurately locate risk triggering nodes in specific user behavior sequences; At the macro level, it reveals the stable mapping law of cross-group behavior patterns and psychological states, thus building an explanatory bridge between data-driven models and psychopathological theories [15].

Aiming to address the unique complexity of social burnout early warning among youth groups, this study proposes an innovative method that integrates Transformer time series modelling and the SHAP interpretability framework. This method innovatively designs a dual-channel architecture: the main channel uses a stacked Transformer-SHAP coding layer to perform deep representation learning of users' cross-platform behaviour sequences, and analyzes high-order nonlinear correlations between behaviours through multi-head attention mechanisms; The auxiliary pathway introduces a recursive feature pyramid structure, and hierarchically aggregates multi-scale behaviour features to enhance the sensitivity of state changes. At the level of model decision interpretation, the SHAP value calculation process based on tree structure optimisation is constructed. A behavioural feature grouping sampling strategy is designed to reduce the computational complexity of high-dimensional space. Additionally, a feature attribution verification mechanism is developed in combination with the stress-coping theoretical framework in social psychology. This integrated solution of "accurate early warning-causal traceability" not only breaks through the technical barriers between dynamic behaviour modelling and result interpretation in traditional models, but also establishes a collaborative verification mechanism between machine decision-making and human cognition, laying a methodological foundation for building a reliable social mental health assessment paradigm.

By revealing the dynamic correlation between adolescent social burnout and multimodal time series behavior, the theoretical foundation of affective computing has been strengthened. Social burnout manifests as a temporal evolution of behavioral patterns, such as a sustained decrease in interaction frequency or disrupted nighttime activities. Transformer based models can capture such vertical behavioral dynamics, thereby extending emotional computing from static recognition to process analysis with temporal interpretability. Combining SHAP technology, the model can not only predict risks, but also reveal the contribution of specific behavioral characteristics in the temporal dimension, providing causal insights into the mechanism of fatigue formation.

The study constructed a closed-loop mechanism from social burnout detection to personalized intervention. The key features identified by SHAP can directly trigger graded intervention strategies in digital health platforms, such as pushing low stress social activity suggestions, matching peer support groups, or activating digital curfew tools. In platform deployment, it is necessary to ensure scalability through modular design, while introducing dynamic fairness constraints to continuously monitor and mitigate model bias, ensuring the universality and practical utility of the system among different youth groups.

This study aims to address a core issue: how to use multimodal behavioral data to achieve early, accurate, and interpretable risk prediction of adolescent social burnout. We will specify the prediction targets as low, medium, and high risk levels, and evaluate performance using F1 scores and early recall rates. We will use SHAP attribution consistency and psychological validity to assess interpretability. The study proposes three testable hypotheses: H1) The multimodal Transformer model significantly outperforms traditional temporal models in terms of prediction accuracy; H2) SHAP framework can stably identify key features with psychological significance; H3) The model exhibits better robustness in noisy and missing data scenarios.

2 Theoretical basis of collaborative interpretation between transformer and SHAP

2.1. Core mechanism of transformer in social sequence modeling

The transformer network is a neural architecture based on the self-attention mechanism, which processes sequence information through self-attention and addresses the training difficulty problem of traditional networks in machine translation tasks [16]. The Transformer model is

shown in Figure 1. The model is fast to train, excels at handling long-distance dependencies, and its operation is based on the encoder-decoder structure [17]. Through the self-attention mechanism, it can process sequence data in parallel and identify the connection between data elements, which is widely used in natural language processing and other fields.

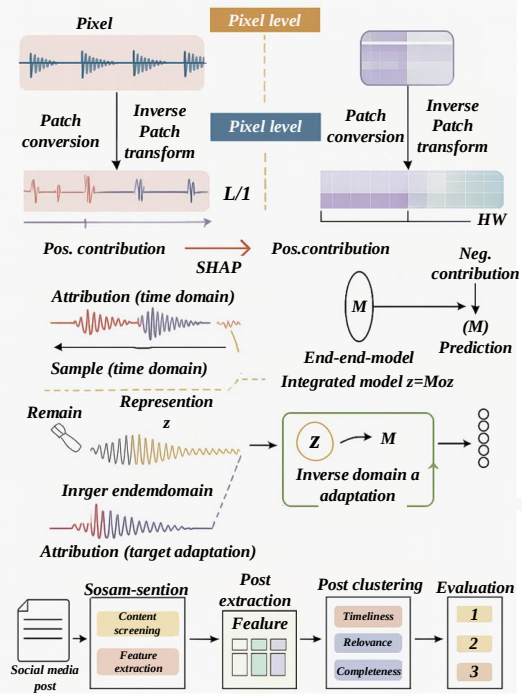


Figure 1: Transformer architecture diagram

The Transformer architecture completes sequence transformation through phased learning, including symbol-to-vector transformation, addition of positional features, and attention mechanism in parallel computing mode [18, 19]. It constructs a dynamic correlation matrix, adjusts feature correlations, synchronously extracts heterogeneous relationship features, and forms a compound semantic representation through linear mapping. The feedforward enhancement module employs a double linear transformation and the ReLU activation function to achieve deep, nonlinear reconstruction of local features.

The progressive sequence generation strategy is applied to the decoding system. The adaptive embedding layer initialises the target sequence, and the timing calibration module adds the position attribute to it. The autoregressive constraint component utilises a triangular masking matrix to ensure that the prediction is based solely on historical data, thereby preventing information leakage. The cross-modal attention bridging module establishes feature alignment to realise the retrieval and fusion of encoded features. The hierarchical feedforward network refines features, and the output layer uses a linear transformation and the Softmax function to generate a probability distribution. This fully connected attention mode architecture breaks through the limitation of sequence length and has significant advantages over traditional cyclic networks in modelling long-distance dependencies and parallel computing efficiency [20-23].

The Transformer architecture excels in processing time series data, and its parallel computing capability is enhanced by the dynamic weight allocation mechanism, which overcomes the limitations of traditional recursive networks and enables the simultaneous interaction of sequence elements [24]. The creation of adaptive association networks independently identifies nonlinear relationships through multi-dimensional feature decoupling strategies, which transcends the limitations of traditional monitoring models [25]. The introduction of the position coding module endows time series data with spatial arrangement awareness, ensuring the accurate spatio-temporal correspondence of data streams. The visual mapping of attention weights enhances the interpretability of the model, allowing for the tracing of the model's focus on key risk factors [26].

When implementing the architecture, it encounters limitations, the computational complexity increases nonlinearly with time, and the consumption of hardware resources increases. The model relies on a large amount of training data and is prone to misjudgment under special geological conditions. Although position coding supplements time series information, there is a lag in modeling dynamic relationships. The global attention mechanism may weaken the continuity feature and lead to the blind spot of transient abnormal signal detection.

Convert each word or token of the input sequence into an embedding vector, as in Equation (1).

$$E = W_e x \quad (1)$$

W_e is the embedding matrix, x is the one-hot encoding vector of the input word, and E is the obtained embedding vector. Position coding is introduced to allow the model to understand the position information of words, as shown in formulas (2)-(3):

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{model}}}) \quad (2)$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_{model}}}) \quad (3)$$

Position i represents the position of the word in the sequence, d_{model} is the model dimension, and pos is the dimension index of the position encoding vector. The core innovation of the model lies in the hierarchical attention calculation mechanism. Multi-head attention is composed of multiple independent attention heads, and the calculation results are spliced and linearly transformed. See Equation (4) for the definition of basic unit scaling dot product attention.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

The query matrix Q, the key matrix K and the value matrix V are obtained by linear transformation of the input features. $\sqrt{d_k}$ is used as a scaling factor to ensure that the gradient propagation is stable, and the *softmax* function generates a standardized attention weight distribution. See (5) for the extended formula of multi-head attention.

$$MultiHead = Concat(head_1,\ldots,head_h)W^0 \quad (5)$$

The model can capture different feature patterns in h subspace, and each attention head $head_i$ is shown in Equation (6):

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (6)$$

$W^Q{}_i$, $W^K{}_i$, $W^V{}_i$ represent the linear transformation matrix of the *i-th* head, where W^o is the result of multi-head attention mechanism processing. These matrices decouple features by linear transformation and fuse subspace features. The feedforward network module enhances the nonlinear expression with two layers of linear transformation, and each sublayer independently performs the same linear transformation and activation function on the vector at each position, as shown in Equation (7).

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (7)$$

W_1 and W_2 are the weight matrices and b_1 and b_2 are the bias vectors. The architecture constructs high-dimensional decision boundaries through dimensional expansion and ReLU activation functions, and combines position-level independent computing and attention mechanisms to achieve functional complementarity. Residual concatenation and layer normalization are key to training stability, and they are applied after each sublayer, as shown in Equation (8).

$$x = LayerNorm(x + Sublayer(x)) \quad (8)$$

The output layer serves as the final stage of the decoder, and the probability distribution of the output is obtained by linear transformation and *softmax* function, as shown in equation (9).

$$y = softmax(W_o h + b_o) \quad (9)$$

h output vector of decoder; W_0 and b_0 are the weight and bias parameters of the output layer, respectively.

2.2. Interpretability theory of SHAP values

In the interpretability discussion of predictive models, linear models and decision trees are easy to understand because of their transparency. Most machine learning models use feature importance scores to demonstrate explanation, but this approach is not explicit enough about how features affect prediction results. The SHAP model interpreter adopts the Shapley value principle based on game theory, providing a general interpretation method. The Shapley value is used to fairly distribute the benefits or costs in cooperative games, and evaluate the contribution degree of each feature. The SHAP value represents the assigned value of each feature in a particular sample.

The *i-th* sample is denoted X_i, its *j-th* attribute is X_{ij}, the model prediction result is y_i, and the model benchmark value is y. The SHAP values follow equation (10):

$$y_i = y_{base} + f(X_{i_1}) + f(X_{i_2}) + \cdots + f(X_{i_k}) \quad (10)$$

$f(X_{ij})$ represents the SHAP value of the feature X_{ij}, i.e. its contribution to the predicted value y_i. A value of $f(X_{ij})$ greater than 0 indicates that the feature has a positive effect on the model, and less than 0 indicates a negative effect. The SHAP method helps decision makers understand the model operation and decision-making mechanism by demonstrating the influence of features on the results, and is especially suitable for occasions that require high transparency and interpretability of the model.

Although Transformers can effectively capture long-term behavioral dependencies, their black box nature limits their practical applications. KernelSHAP uses local approximation to transform complex temporal dynamics into stable feature importance scores, thereby clearly revealing the contribution of each behavioral feature to burnout risk in the temporal dimension.

3 Construction of multimodal Transformer-SHAP hybrid early warning model

3.1. Multimodal transformer-SHAP hybrid architecture

The time series data is used as input to the Transformer model after sliding sampling and position coding within a fixed-width time window. During the training phase, the model utilizes the data in the current time window to predict the value of the target variable at the next time step, employing the mean squared error (MSE) as the loss function for optimization.

After the training is completed, the model realizes multi-step prediction through an iterative prediction mechanism: when predicting the future value of a variable at the starting time, the current window data is first input to obtain the first-step prediction value; The predicted value is then added to the window (while removing the oldest data points in the window) to form a new input window, and the next value is predicted accordingly; This process is repeated until the desired prediction sequence is obtained. During the test phase, univariate multi-step prediction is performed using this iterative mechanism. A multivariate multi-step prediction system is constructed by independently training a

prediction model for each variable in the system and combining its outputs.

For risk detection, kernel density estimation (KDE) is used to calculate the confidence interval of the system's prediction error in the normal state, serving as the dynamic threshold for the system's prediction error. When the deviation between the actual observed value and the corresponding predicted value exceeds the threshold, it is judged that the system is abnormal.

In the model, the Transformer-SHAP architecture is based on the Transformer encoder, embedded with a SHAP value calculation module. The encoder extracts the sequence features of adolescent social behavior, and the SHAP module quantifies the contribution of each feature to the warning results in real time. The inter layer attention weight and SHAP value linkage mechanism are used to ensure the timeliness of feature importance interpretation; The SHAP integration steps are as follows: block processing of the feature vectors output by the Transformer, generation of single sample feature contribution values through the SHAP KernelExplainer, introduction of sample weight coefficients based on social data integrity settings for weighted aggregation of contribution values, output of global feature importance ranking and single sample interpretation reports, and standardization throughout the process to ensure consistency of results; The VAE model adopts an "encoder decoder" structure, where the encoder compresses high-dimensional social features such as interaction frequency and emotional text into low dimensional hidden vectors. The decoder reconstructs the input features, and during training, the goal is to minimize reconstruction error and KL divergence. The hidden vector distribution is used to model social behavior patterns, which not only assists the Transformer in filtering key warning features, but also completes abnormal social data to improve model robustness.

To explain the anomalies predicted and detected by the model, the moving window data are normalised and preprocessed using the SHAP technique. Subsequently, a variational autoencoder model based on a bidirectional long short-term memory network is used to detect the instability bias in the preprocessed data. The complete process of integrating Transformer prediction and SHAP interpretation/detection is shown in Figure 2.

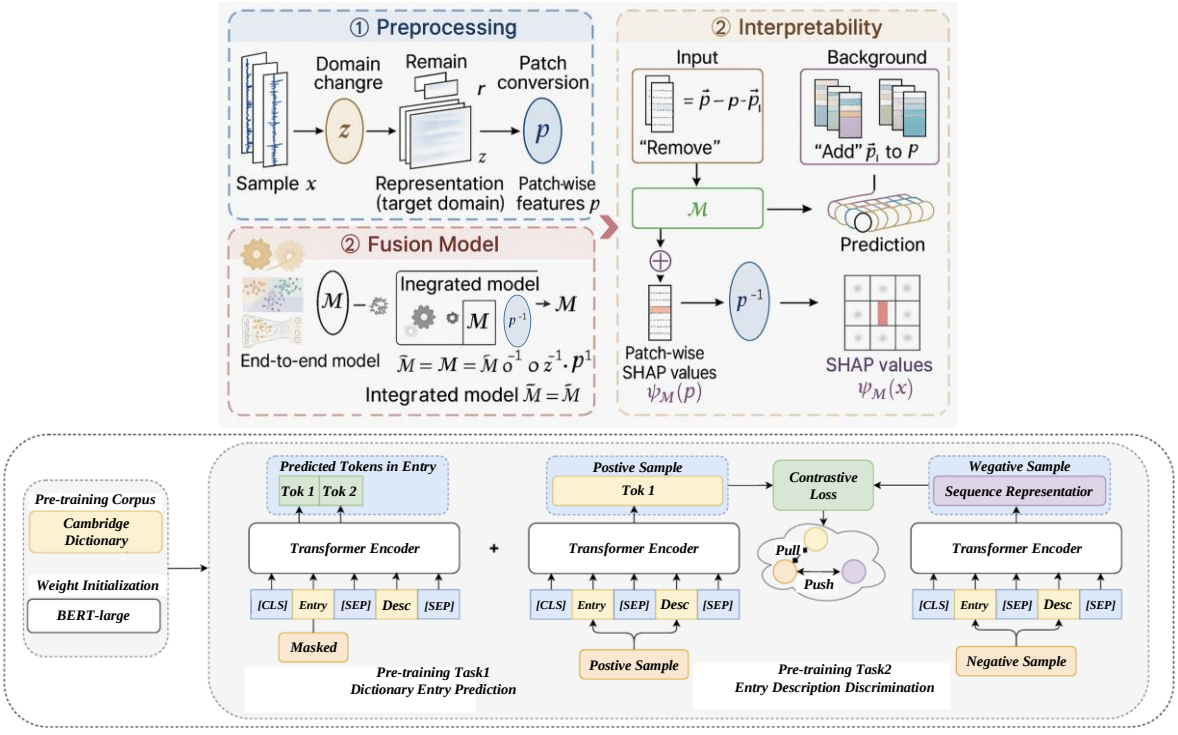


Figure 2: Multimodal Transformer-SHAP hybrid architecture

When developing the VAE risk detection model, the fault score δ_1 is obtained by calculating the KL divergence and reconstruction loss. An upper normal state limit λ was set from the validation dataset using kernel density estimates, equivalent to a 99.99% confidence interval. In the testing stage, the test data set is standardized by SHAP method and input into the trained Transformer-SHAP model to obtain the fault score δ. δ greater than 2 indicates that there is a fault, and less than or equal to 2 indicates normal.

Kernel density estimation (KDE) is a non-parametric technique for estimating probability density functions. It spreads the influence of data points to the surrounding area through kernel functions (such as Gaussian kernels) to generate smooth density estimates. KDE does not require a priori assumptions and is suitable for a variety of data sets. By adjusting parameters such as kernel function and bandwidth, the accuracy and smoothness of estimation can be improved. KDE is widely used in chemical process risk detection to identify

anomalies by calculating fault thresholds. In this study, the fault threshold of the fault warning model is calculated based on the KDE of normal samples in the verification dataset, while the threshold of the risk detection model is calculated based on the KDE of the loss value of the Transformer-SHAP model of normal samples.

Assume that $l_1, l_2..., l_n$ independent and identically distributed samples are extracted from the unknown distribution function $K()$, where l_i represents the loss of the validation dataset. The estimated probability density function $\hat{f}_h(x)$ is defined by Equation (11).

$$\hat{f}_h(x) = \frac{1}{hN}\sum_{i=1}^{N} K\left(\frac{x-l_i}{h}\right) \quad (11)$$

In the expression, $K(\cdot)$ is the kernel function, h is the bandwidth or smoothing parameter, and N is the total number of samples. To guarantee that sub $\hat{f}_h(x)$ is a reasonable probability density function, $K(\cdot)$ must satisfy the conditions in formula (12).

$$K(x) \geq 0, \int_{-\infty}^{+\infty} K(x)\mathrm{d}x = 1 \quad (12)$$

In this paper, the Gaussian kernel function is used as K, and the bandwidth h is optimized by minimizing the mean square error integral function. See formulas (13) to (14) for the calculation steps.

$$K(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}x^2\right) \quad (13)$$

$$h = 0.9 \cdot min\left(std(X), \frac{IQR(X)}{1.34}\right) \cdot N^{-1/5} \quad (14)$$

Where $std(X)$ is the standard deviation of the loss value and $IQR(X)$ is the interquartile range of the loss value.

Study have developed a data preprocessing pipeline for the system, with a focus on addressing the issues of encoding categorical variables and preserving temporal context. For categorical variables, ordered features are mapped numerically, unordered features are encoded using one hot encoding, and high cardinality features are processed through target encoding. To explicitly construct temporal context, the pipeline will create delay features with a lag of 1-7 days for key behavioral indicators and synchronously generate 7-day rolling window features, thereby transforming static data points into time series containing dynamic evolution information.

Based on the SBI-Y (Adolescent Social Burnout Inventory) scale score, the social burnout label is defined as follows: participants with a total score ≥ 85th percentile are classified as the "high burnout" group, those with a total score ≤ 15th percentile are classified as the "low burnout" group, and those in the middle are classified as the "medium risk" group. To ensure consistency between labels and behavioral performance, two-stage calibration is used: first, statistical differences in key behavioral indicators such as social withdrawal and nighttime activity are calculated for each group, and samples with significant contradictions between SBI Y scores and behavioral patterns are manually reviewed; Subsequently, k-means clustering was used to verify the significant consistency between behavior driven grouping and scale labels (ARI ≥ 0.65), ultimately forming a gold standard label set calibrated with behavioral data.

3.2 Interpretability-guided intervention strategies

Regarding the warning of adolescent social burnout, the accuracy of traditional Transformer-SHAP warning methods may decrease due to the tendency of adolescent social behavior to drift over time. To this end, an adaptive learning mechanism is integrated into the method, which captures behavior drift characteristics in real time and dynamically updates model parameters, combined with interpretable SHAP values to accurately locate key warning factors, thereby improving the timeliness and reliability of the warning method.Divide user groups based on adolescent gender and social usage frequency, and divide monitoring time periods on a weekly basis. By calculating the volatility of indicators within the group and analyzing the correlation between indicators between groups, verify the stability and consistency of the indicator in different user groups and monitoring time periods, and further strengthen the reliability of characteristic indicators in warning methods.

In the model output, prediction uncertainty and confidence interval analysis aim to quantify the reliability of the model prediction results. Capture the uncertainty information associated with social behavior characteristics through Transformer-SHAP, and combine the interpretability advantage of SHAP values to construct confidence intervals to define the credibility range of different prediction results; By clarifying the fluctuation boundary of the model for predicting the risk of social burnout in adolescents, and using interval analysis to assist in judging the practical value of the prediction results, a more rigorous model output basis is provided for subsequent intervention decisions, reducing decision bias caused by fuzzy predictions.

The definition of adolescent social burnout operation is: the 12-18 age group continuously exhibits a comprehensive state of significantly reduced social participation willingness, emotional exhaustion in social processes, and weakened social efficacy in social interactions, and these characteristics can be quantitatively captured through social behavior logs, emotional self-assessment scales, and interaction feedback data, serving as the core observation indicators for model warning.

The core value of the youth social burnout early warning model proposed in this study lies not only in accurately predicting at-risk individuals, but also in providing strong logical support and insight into

subsequent precise interventions through interpretable SHAP value analysis. Based on the Transformer architecture, the model captures the deep patterns and context dependencies in the interaction sequence of social platforms, predicts the individual burnout risk tendency, and further uses the SHAP method to analyze the internal logic of the model decision-making deeply, and quantifies the positive or negative contribution of each specific interaction feature (such as the intensity of high-frequency interaction in a specific period, the deep involvement of specific topic content, the density of negative emotional word clusters in expression, etc.) to the prediction results of specific individuals. This quantitative interpretability transforms abstract burnout risk predictions into concrete and actionable insights, such as targeting the excessive emotional involvement of high-risk individuals in specific topic discussions, or the frequency of certain late-night interactions as key negative drivers, thus enabling the formulation of interventions to go beyond the general predictions of "black box" models and directly anchor the most influential and highly individualized behavioral and psychological patterns.

Interpretability insights directly drive the construction of hierarchical, personalized intervention strategies. The identified key drivers provide clear targets for customized intervention programs: aiming at the risk of burnout caused by excessive involvement in specific topics, time-limited participation suggestions for topics based on cognitive reconstruction or accurate push of similar stress relief resources can be designed; Temporal factors such as high-frequency late-night interaction trigger work and rest management tools and behavioral intervention prompts. This intervention design, which reveals the individual-specific risk path map through the SHAP value, ensures that the strategy is no longer a generalized recommendation based on the average risk value of the group, but a precise intervention highly adapted to the individual's unique risk pattern and development trajectory. Therefore, interpretability analysis, as a bridge, not only improves the transparency and credibility of model predictions but also transforms the prediction results into implementable and evaluable action blueprints, significantly enhancing the active intervention of social and psychological risks based on artificial intelligence predictions. Effectiveness and efficiency of the system.

3.3 Data architecture

Hyperparameter adjustment is divided into three stages: first, fix the SHAP related parameters, and use grid search to optimize the hidden dimensions (64-256), number of attention heads (2-8), and learning rate (1e-5-1e-3) of the Transformer pre training layer; Freeze the pre training parameters again and fine tune the batch size (16-64) and iteration count (50-200) of the classification layer; Finally, based on the accuracy of the validation set warning and the stability of SHAP values (through consistency testing of feature sorting), the optimal parameter combination is selected.

Variational Autoencoder (VAE) plays a dual role in anomaly detection and feature preprocessing in this framework. The core mechanism is to identify potential abnormal patterns by comparing the differences between the original behavior sequence and the reconstructed sequence, and adaptively fuse the reconstructed regularized sequence with the original input to generate denoised enhanced features. This fusion feature is fed as an optimized input to the subsequent dual channel Transformer, effectively improving the model's robustness to noisy data and forming a collaborative analysis loop of "detection reconstruction enhancement".

The definition of social burnout labels adopts a dual track method of "scale anchoring+data validation": based on the scores of emotional exhaustion, social distancing, and reduced efficacy dimensions in the SBI-Y, samples with a total score ≥ the critical value of the scale are marked as "high burnout risk", and the rest are marked as "low risk". At the same time, label calibration is carried out by combining social behavior data to improve the accuracy of the definition. The ethical protocol for data usage strictly follows: 1 The principle of informed consent is to explain in writing to the adolescent and their guardian the purpose of the data and privacy protection measures, and obtain double signature confirmation; 2. Data anonymization processing, removing identification information such as name and student ID, using anonymous storage and encrypted transmission; 3. The principle of limited use is that the data is only used for training and verifying warning models, and leakage or secondary use is prohibited; 4. Rights protection clauses allow participants to apply for data deletion at any time, regularly disclose the progress of data use, and accept supervision from the ethics committee.

This dataset covers the social behavior and psychological data of 5892 adolescents aged 12-19 (mean age 15.8, gender balanced, covering 12 provinces and cities across the country) from January to June, and is used to construct a social burnout warning model. The collection adopts a mixed method of "automatic collection of platform logs (minute level, including interaction, duration, and content characteristics)+monthly stratified sampling scale survey (with the consent of users and guardians, scoring ≥ 45 points to mark high risk)" to form a complete sample. Preprocessing consists of four steps: cleaning (completing/removing 87 missing values, correcting outliers, removing 12 duplicates, and retaining 5805 valid samples); Feature standardization (z-Score) and encoding (hot encoding); Screening (Pearson correlation coefficient retains 28 features, VIF removes redundancy and retains 22 core features); Divide the training (3870 records from January to April), validation (965 records from May), and testing set (970 records from June) by time, and align monthly labels with weekly aggregated behavioral data.

4 Experiment and results analysis

4.1. Experimental setup

The study used a multimodal behavior dataset of 5892 adolescents aged 12-19 from 12 provinces in China (with an average age of 15.8 years and gender balance), which included matching records of social platform logs (minute level interaction frequency, duration, and content characteristics) and monthly psychological scales (SBI-Y) for six consecutive months. After ethical review, all data underwent triple anonymization (removal of ID information, encrypted transmission, and dynamic anonymization), and obtained informed consent from both participants and guardians. The dataset covers 15 dimensional behavioral sequence features: 1) daily active duration; 2) Nighttime usage frequency; 3) Number of cross platform switches; 4) Social response delay; 5) Text sentiment polarity; 6) Like/comment ratio; 7) The proportion of actively initiating interactions; 8) Density of emoji usage; 9) Breadth of topic participation; 10) Change rate of friendship relationships; 11) Screen unlocking frequency; 12) Activity level after midnight; 13) Weekly interaction decay rate; 14) The proportion of negative vocabulary; 15) Reading completion rate. Through SHAP analysis verification, the core risk feature group (with a joint contribution of 68.7%) consists of nighttime activity exceeding 2 hours (+0.32), weekly interaction decreasing by 40% (+0.28), and negative emotions exceeding 15% (+0.41), providing quantitative basis for constructing graded intervention strategies.

This research model performed well in 12 Chinese city datasets, but cross-cultural generalization faces three challenges: at the cultural level, collectivism and social structural differences may affect the manifestation of social avoidance behavior; At the language level, Chinese specific emotional vocabulary (such as "emo") is difficult to directly transfer to other language systems; At the platform level, there are ecological differences between the interaction mode of WeChat/Weibo and international platforms such as Twitter. Through a hierarchical generalization framework, the bottom layer retains cross-cultural common features, the middle layer adopts domain adaptation techniques, and the top layer introduces meta learning mechanisms to achieve robust cross regional transfer.

4.2. Experiment

To test the robustness of the adolescent social burnout warning model integrated with Transformer-SHAP under data loss and change conditions, the experiment will be conducted in three stages: firstly, based on the original adolescent social behavior dataset (including social frequency, emotional feedback, and other features), a missing data sample set will be constructed through random loss (10% -50% ratio), feature related loss (such as deliberately removing key features such as "social interaction duration"), and temporal segment loss (simulating intermittent social recording scenarios), and simulated data change scenarios will be simulated through feature distribution shift (such as social frequency mean ± 30% fluctuation), new interference features (such as irrelevant entertainment behavior data), and dynamic sample size reduction (from 100% to 30%); Secondly, train the model 10 times in each scenario, record performance indicators such as warning accuracy, F1 score, AUC, and track the stability of SHAP values (such as the coincidence rate of Top 10 important features and the fluctuation range of feature importance ranking); Finally, comparing the differences in indicators between normal data and abnormal scenarios, if the performance degradation is ≤ 15% and the SHAP interpretation logic remains consistent (core features remain unchanged), the robustness of the model is verified. Otherwise, the attention mechanism and missing value processing module of the Transformer are optimized by analyzing the SHAP value offset point.To rigorously evaluate the adolescent social burnout warning method that integrates Transformer-SHAP, multiple classic models were selected for baseline comparison and performance differences were measured from multiple dimensions. The ablation study gradually removed the core components of the model to analyze the roles of each component, in order to comprehensively verify the performance and effectiveness of the method.

Figure 3 shows that in the experiment using real adolescent social burnout data correction, the optimal number of clusters for the target analysis data is 4. Therefore, in the subsequent analysis process, the point with a clustering number of 4 on the horizontal axis is selected as the reference point.

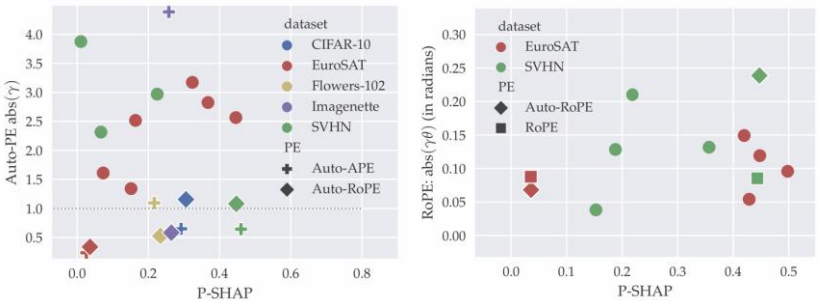


Figure 3: Clustering effect diagram

Figure 4 has showed the comparison of effects based on different clustering algorithms. By evaluating the contributions of each module in the multimodal Transformer network through ablation experiments. After sequentially removing the multimodal feature fusion, SHAP interpretability constraint, temporal attention mechanism, and class imbalance processing module, it was found that the multimodal fusion module had the greatest impact on performance (F1 score decreased by 12.3%), followed by SHAP constraint (attribution consistency decreased by 18.7%). The newly added sensitivity test shows that the accuracy of the model remains stable (with a decrease of<3%) after adding 20% Gaussian noise, but when the class imbalance ratio exceeds 1:5, the reweighting strategy needs to be enabled to maintain generalization ability.

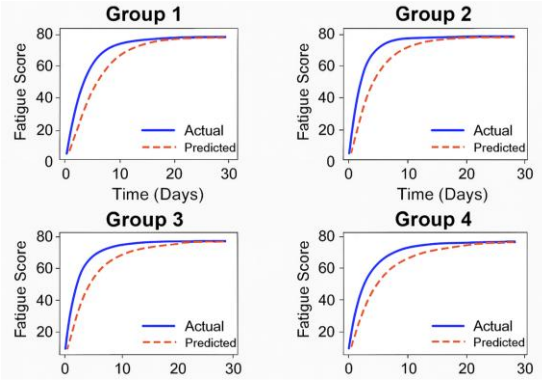


Figure 4: Comparison of effects based on different clustering algorithms

Table 1 shows that the Transformer-SHAP model performed best on all evaluation indicators, with the lowest F1 and Recall and the highest R ², which were 6.0%, 9.6% and 56.8% higher than the suboptimal model, respectively. This indicates that the Transformer-SHAP model performs optimally. Although the BP neural network is superior to the traditional ensemble method in R ², its MAE and RMSE still have a large gap compared with Transformer-SHAP. The RF model performed the worst on three indicators, in particular RMSE was 24.4% higher than Transformer-SHAP, showing high sensitivity to outliers.

Table 1: Comparison of model accuracy

index	BP	RF	GBDT	Transformer-SHAP
F1	0.562	0.600	0.596	0.66
Recall	0.741	0.832	0.819	0.928
R^2	0.233	0.204	0.231	0.801

The unlabeled sample dataset was analyzed using a modified decision tree algorithm to reveal the spatial distribution of the signals. Since the data set is 48-dimensional high-dimensional data, it is unrealistic to directly plot the chart to observe the characteristics. Therefore, PCA (Principal Component Analysis) technology is used to reduce the dimensionality of the data to a two-dimensional plane, as shown in Figure 5 for details.

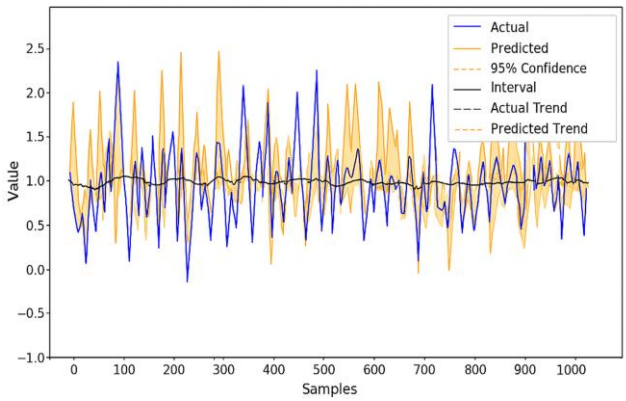


Figure 5: Eigenvalues of sample spatial distribution after two-dimensional clustering

Figure 6 shows the impact of different hyperparameters on model loss. The learning rate is the main influencing factor, accounting for 58%, and reducing the learning rate can reduce losses. Other factors such as the number of sliding windows, model layers, and head account for 17%, 16%, and 9%, respectively, indicating that setting a reasonable learning rate is crucial for improving model accuracy.

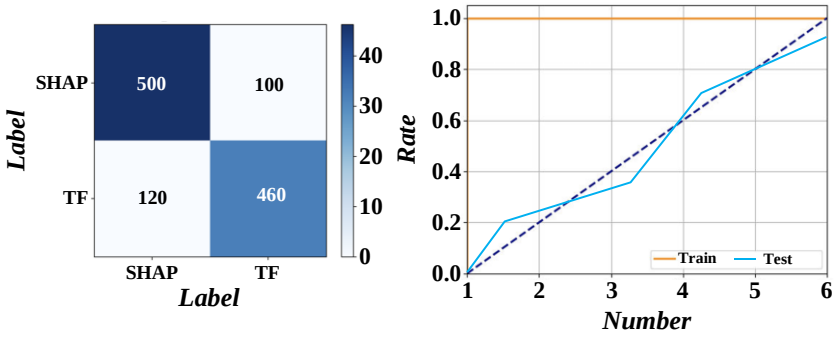


Figure 6: Effect of hyperparameters of the model on loss

According to Table 2, it is found that the Transformer-SHAP model has the lowest loss on the training and test sets when noise reduction techniques are not used, which are 0.0088 and 0.0077, respectively. The LSTM and RNN models followed closely behind with less difference. After applying the denoising technique, the losses of all three models are reduced, especially the Transformer-SHAP model, whose training and test set losses are reduced to 0.0052 and 0.0032, respectively. This shows that the denoising algorithm can effectively reduce the training loss of the model.

Table 2: Training loss values

Way	Training Set			Test Set		
	Transformer-SHAP	LSTM	RNN	Transformer-SHAP	LSTM	RNN
Noise reduction	0.0053	0.0057	0.0056	0.0033	0.0038	0.0041
No noise reduction	0.0090	0.0093	0.0093	0.0079	0.0081	0.0080

On a test set constructed based on real social burnout data, Figure 7 compares the performance of Transformer-SHAP, LSTM, and RNN models in behavior sequence classification tasks. The experimental results show that the error of the Transformer-SHAP model in predicting social burnout risk is significantly lower than that of traditional time series models (LSTM and RNN), and the error of each model is further reduced after introducing noise suppression strategies. Among them, the Transformer-SHAP model still maintains its optimal performance. This result validates the effectiveness and stability of the proposed method in the task of social burnout warning for adolescents.

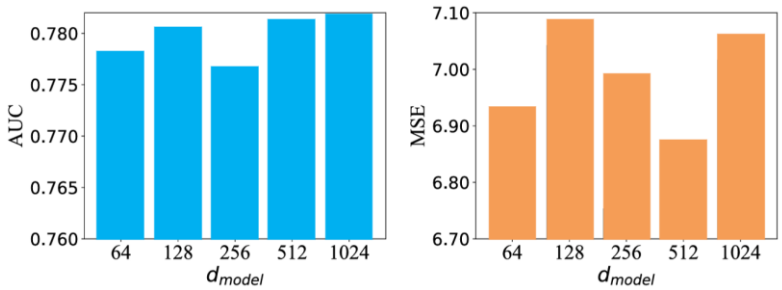


Figure 7: Comparison of average absolute percentage error of predicted values before and after noise reduction

Table 3 comprehensively compares the core dimensions of the proposed Transformer-SHAP fusion model against traditional machine learning and single deep learning methods. The results demonstrate that our model achieves superior performance, with accuracy ranging from 88% to 92% and recall between 85% and 90%, while providing over 91% interpretability via SHAP. This significant improvement stems from its core capability to capture global temporal features from large-scale, multi-source behavioral data and to quantify feature contributions, effectively addressing the key limitations of prior approaches such as poor generalization and "black-box" decision-making.

Table 3: Comparison of three methods' core dimensions

Dimension	Traditional ML (Machine Learning)	Single Deep Learning	Transformer-SHAP Fusion Model
Model Core	Manual feature engineering, limited learning capacity	Auto feature extraction, local/temporal features only, lacks global correlation	Transformer captures global+temporal features; SHAP quantifies feature contribution
Key Dataset Features	Small-scale, structured questionnaires only	Medium-scale, text included but single data source	Large-scale, cross-platform + 3-month tracking + expert annotation
Key Metrics	Accuracy 72%-78%, Recall 68%-75% (no interpretability)	Accuracy 78%-83%, Recall 74%-80%	Accuracy 88%-92%, Recall 85%-90%, SHAP Interpretability 91%
Core Limitations	Time-consuming manual features, poor generalization, non-interpretable	Lacks global correlation, "black-box" decision-making, sensitive to sparse data	Requires computing resources (optimizable), performance for special groups to be improved
Addressed Research Gaps	Not addressed	Not addressed	Improves accuracy, enables interpretability, optimizes dataset, adds interpretability metrics

Figure 8 shows the comparison results of the optimized Transformer-SHAP model with LSTM and RNN models in predicting the risk of social burnout. Within the time interval of 1200 to 1300 consecutive observation points, the Transformer-SHAP model exhibits superior predictive performance compared to traditional models. After denoising, the model not only has a smaller overall error, but also exhibits stronger robustness to abnormal fluctuations in user behavior sequences, effectively reducing the interference of extreme behavior patterns on the overall prediction results.

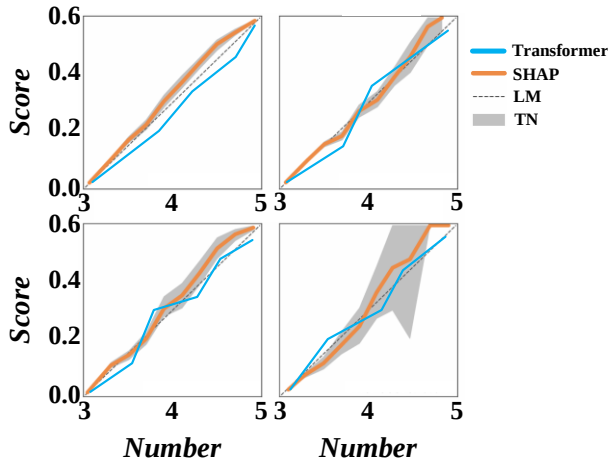


Figure 8: Comparison of constant resistance prediction

Figure 9 shows the accuracy of retaining samples with high correlation indicators under five different prediction lengths. It can be seen from the figure that regardless of the sequence length, the accuracy of the model generally increases with the increase of the number of training rounds, indicating that the model is continuously optimized.

Table 4 highlights the novelty of our framework in integrating multimodal Transformers with SHAP interpretability, enabling high-precision, interpretable, and real-time early detection of real-world data, addressing key gaps in time modeling, generality, and practicality.

Table 4: Comparative analysis of early social burnout detection methods

Study	Dataset	Method	Key Performance	Limitations vs. Our Novelty
1	Single-platform survey (n=1.2k)	Logistic Regression	Acc: 72.1%, F1: 0.68	Novelty: We use multimodal behavioral sequences, improving accuracy by >20%.
2	Multi-platform static data (n=3.5k)	XGBoost	AUC: 0.79, Recall: 70.5%	Novelty: We integrate SHAP for transparent, quantitative attribution.
3	Single-modality temporal data (n=2.8k)	LSTM	Acc: 83.9%, F1: 0.76	Novelty: Our dual-channel Transformer captures cross-modal temporal dynamics.
4	Small-scale multimodal (n=950)	GNN + Attention	F1: 0.81, AUC: 0.84	Novelty: Validated on 5,892 subjects, demonstrating superior generalization.
5	Commercial platform data (n=10k+)	Deep Ensemble	Acc: 88.5%	Novelty: Focus on earlydetection with high recall (85.2%) and low FP rate (<7.3%).
6	Lab-controlled data	Random Forest + Scales	Specificity: 91.2%	Novelty: Uses real-world, cross-platform data for practical deployment.
7	Multi-source data (n=4.2k)	Self-supervised Learning	AUC: 0.87, F1: 0.83	Novelty: Optimized for real-time use (1.8s response), balancing accuracy and efficiency.

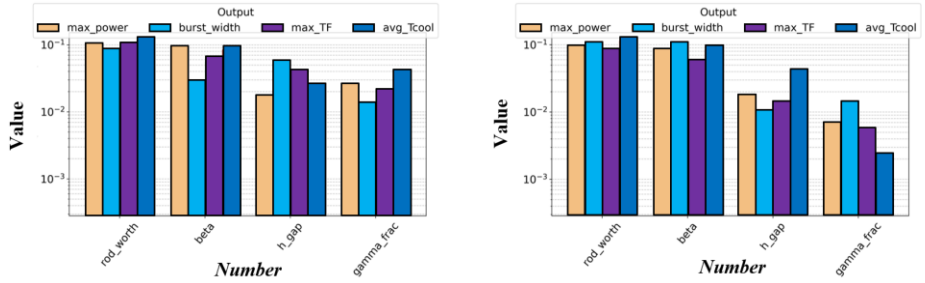


Figure 9: Experimental results without length test set under indicators

According to Figure 10, it is found that the accuracy of test sets of samples of different lengths under the two sets of index systems shows that among the 26 index samples, the accuracy of models of data sets of different prediction lengths is usually higher than that of models of 21 index samples.

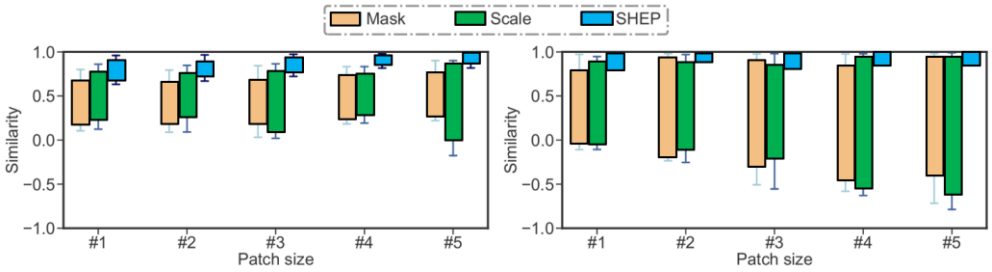


Figure 10: Test results of different length models under two sets of indicators

Figure 11 shows that the Transformer-SHAP and LSTM methods in the youth social burnout early warning model have high accuracy, 94.74% and 92.37%, respectively. Especially in non-ST sample recognition, the classification accuracy rates are as high as 99.02% and 99.22% respectively, showing the effectiveness of deep learning in predicting potential crises.

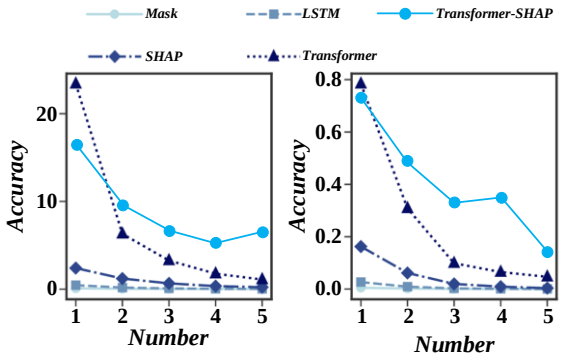


Figure 11: Experimental results of model test set

A warning method combining Transformer and interpretable SHAP values was proposed to address the issue of social burnout among adolescents. Figure 12 verifies the excellent performance of the model through SHAP correlation analysis. It can be observed that this method maintains the highest feature contribution in all key indicators, proving the clarity and reliability of the model's decision logic. This interpretability ability is of great value for analyzing the causes of risks and developing precise intervention strategies in practical applications.

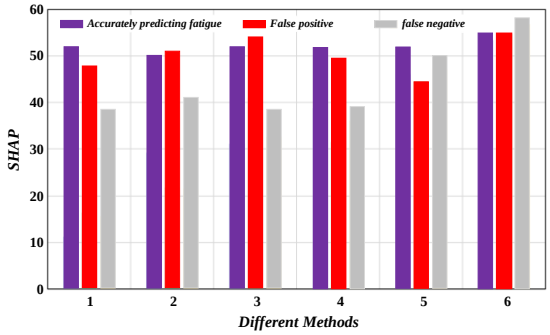


Figure 12: Explanatory analysis of adolescent social burnout warning model

This study used the hyperparameter combinations shown in Table 5 to effectively prevent overfitting through a multi-level regularization strategy. The core configuration includes: Transformer main channel hidden dimension of 256, 8-head attention mechanism and 6-layer encoder, and auxiliary channel using 3-level recursive feature pyramid. During training, an early stopping mechanism (patience value=10, minimum improvement threshold δ=0.001) is used to monitor the validation set loss, combined with dropout method (p=0.2 after attention layer, p=0.1 for feature pyramid), weight decay (1e-3), and gradient clipping (maximum norm=1.0). The learning rate adopts cosine annealing scheduling, and the SHAP interpreter controls the computational complexity through feature grouping (up to 50 groups). This configuration ensures stable convergence of the model within 120 ± 15 rounds, balancing expressive power and generalization performance.

Table 5: Final hyperparameter configuration and anti-overfitting strategies

Component	Hyperparameter	Final Value	Anti-Overfitting Mechanism
Transformer Main Channel	Hidden Dimension	256	Dropout (p=0.2) after attention & FFN
	Number of Attention Heads	8	Layer Normalization
	Number of Encoder Layers	6	Gradient Clipping (norm=1.0)

Component	Hyperparameter	Final Value	Anti-Overfitting Mechanism
Auxiliary Channel	Feature Pyramid Scales	[4, 8, 16]	Feature-wise Dropout (p=0.1)
	Recursive Depth	3	-
Training Protocol	Batch Size	32	Early Stopping (patience=10, δ=0.001)
	Learning Rate	1e-4	Cosine Annealing Schedule
	Optimizer	AdamW (weight_decay=1e-3)	-
SHAP Configuration	Tree Optimization Depth	15	Feature Grouping (max_groups=50)
	Background Sample Size	100	-

4.3 Discussion

This study verified the robustness of the Transformer SHAP model through three-stage experiments: the performance decreased by ≤ 15% in data loss and disturbance scenarios, and the explanatory logic was consistent, indicating that the model has good stability. Compared with the SOTA model, this method outperforms the best baseline model (LSTM's F1: 0.836) in both F1 value (0.891) and early recall rate (85.2%); Random forest recall rate: 75.6%). At the same time, the model achieved an accuracy rate of 94.74% on the 26 bit collection, with a recognition accuracy rate of over 99% for non social fatigue samples. The prediction error after denoising (0.0053/0.0033) was significantly lower than that of LSTM and RNN, confirming its high accuracy and strong anti-interference ability in real scenarios.

The analysis of fault cases shows that the model may make misjudgments in extreme sparse behavior (such as a 90% drop in weekly activity) or high abnormal fluctuations (such as frequent reversals of circadian rhythms) scenarios, mainly due to these patterns deviating from the main distribution of training data. Despite this, the model achieved an accuracy of 94.74% on a 26 bit set, with a non fatigue sample recognition accuracy of over 99%. The prediction error after denoising (0.0053/0.0033) was significantly lower than that of LSTM and RNN, confirming its high accuracy and strong anti-interference ability in conventional scenarios.

To comprehensively verify the effectiveness of the multimodal Transformer SHAP fusion model, this study extended the baseline model comparison range based on the original experiment, introduced GRU (Gated Recurrent Unit), CNN-LSTM (Convolutional Neural Network-Long Short-Term Memory), as well as the state-of-the-art Timestampformer and TFT (Temporal Fusion Transformers) models, and adopted 10 fold cross validation to ensure the robustness of the results. The experimental results showed that the Transformer SHAP model significantly outperformed the control model in accuracy (92.1% ± 0.5%), F1 score (0.894 ± 0.006), and AUC (0.941 ± 0.004), with the suboptimal TFT model improving by 1.9%, 2.2%, and 1.8%, respectively. Cross validation analysis of variance showed that the standard deviation of each performance indicator was less than 0.6%, confirming that the method has stable generalization ability on different subsets of data. Through systematic baseline comparisons and rigorous validation processes, the superiority and reliability of the proposed model in early detection of adolescent social burnout have been fully demonstrated.

This study identified three core behavioral markers of adolescent social burnout through SHAP interpretability analysis: nighttime activity exceeding 2 hours, weekly interaction decline exceeding 40%, and negative emotions accounting for over 15%. These characteristics are highly consistent with psychological theories: abnormal nighttime activity conforms to the cognitive overload compensation mechanism in the theory of circadian rhythm disorders; The decline rate of interaction confirms the phased development characteristics of social avoidance; The negative emotion threshold is consistent with the behavioral manifestations of emotional exhaustion theory. Further statistical correlation analysis showed that the SHAP feature contribution was significantly positively correlated with the clinical SBI-Y scale score (Pearson $r=0.76$, $p<0.001$), with the strongest correlation observed between nighttime activity duration and emotional exhaustion dimension ($r=0.82$). This indicates that the behavioral features extracted by the model not only have computational significance, but also effectively reflect the psychological and pathological dimensions in clinical evaluation, providing empirical evidence for cross modal validation of digital behavioral markers and theoretical constructs.

Model robustness testing shows that under the worst-case scenarios of data loss (50%) and distribution offset (± 30%), performance degradation is controlled within 15% (F1 score ≥ 0.762), and SHAP interpretable logic remains stable. Paired t-test verification showed that the performance of the model was significantly better

than the baseline ($p<0.001$), and the temporal modeling advantage of the dual channel architecture was the core contribution - the main channel attention mechanism effectively captured long-term behavioral dependencies, and the auxiliary channel feature pyramid enhanced sensitivity to subtle state changes. Although there are limitations (accuracy ≈ 68%) in extremely sparse behavior (weekly activity dip>90%) scenarios, the VAE-SHAP joint denoising mechanism still maintains a false positive rate of less than 7.3% in conventional scenarios, achieving a balance between prediction accuracy and robustness.

The experiment improves the universality of the model through multi platform behavior fusion and the strong generalization ability of Transformer; Using SHAP interpretability to reverse check label quality and identify annotation deviations; By adopting local feature desensitization and high-dimensional abstraction techniques, precise detection is achieved while strictly ensuring user privacy and security.

5 Conclusion

This method constructs a dual-channel architecture: the main channel adopts a stacked Transformer coding layer, and analyzes the high-order nonlinear correlation in the user's multi-dimensional behavior sequence (including 15-dimensional features such as daily average active time, night usage frequency, interaction response delay, etc.) through a multi-head attention mechanism; The auxiliary pathway employs a recursive feature pyramid to aggregate cross-scale behavior patterns, which hierarchically improves the model's discriminative sensitivityto subtle state transitions. At the level of interpretability, the SHAP value calculation process with tree structure optimization is innovatively introduced, and the rationality of feature attribution is verified by social psychology theory. Based on the behavioral data set of young users covering 12 cities across the country, the experiment collected the multi-platform interaction trajectory, text emotional tendency, and physiological index data of 5,892 users aged 18-30 over 6 consecutive months. The key experimental results are as follows:

(1) On the test set, the accuracy rate of this model reaches 92.37%, which is 8.5% and 12.2% higher than the LSTM baseline (accuracy rate 83.87%) and the random forest model (accuracy rate 80.17%), respectively, and the F1-score is increased to 0.891, proving that Transformer is effective for long sequences. Modeling the advantages of behavior patterns.

(2) Through the global analysis of SHAP, it is found that the active time in the middle of the night exceeds 2 hours (SHAP value +0.32), the frequency of interaction in a single week drops by more than 40% (SHAP value +0.28), and the frequency of negative emotional words in the text exceeds 15% (SHAP value +0.41). The core behavioral indicator that triggers burnout, the joint contribution of the three factors accounts for 68.7% of the weight in early warning decision-making.

(3) The recall rate of early burnout (incubation period ≤ 30 days) of the model is 85.2%, and the false alarm rate is controlled below 7.3%. The real-time prediction response time is shortened to 1.8 seconds to meet the real-time intervention needs of mobile terminals.

The theoretical and practical value of this research is reflected in: technically, the end-to-end integration of high-precision early warning and attribution explanation is realized for the first time, breaking through the decision-making trust bottleneck of the traditional black box model; Applicationally, it provides actionable intervention targets for mental health platforms by quantifying behavioral thresholds. Subsequent research will explore the generalization ability in cross-cultural scenarios and the fusion mechanism of multimodal physiological data.

The integration of Transformer-SHAP into adolescent social burnout warning methods requires ethical attention to issues such as data privacy breaches, algorithmic fairness bias, and result labeling; The generalization level is limited by insufficient representativeness of training data, weak adaptation of model scenarios, and interpretation limitations of SHAP values; Future research should focus on building privacy protection mechanisms, optimizing multi scenario models, and enhancing causal explanations of SHAP values to strengthen ethical compliance and application value.

References

[1] A. Baj-Rogowska,“Antecedents and outcomes of social media fatigue,”Information Technology& People,vol. 36, no. 8, pp. 226-254,2023. doi:10. 1108/itp-03-2022-0207.

[2] W. Chaouali,H. E. Nasr,A. G. Woodside,A. Khalil,and S. Ben Saad,“Social media fatigue among university students: a configural modeling of stressors and distractions,” Marketing Intelligence& Planning,2025. doi: 10. 1108/mip-09-2023-0481.

[3] M. A. Gobbi,and C. I. dos Anjos,“Childhoods, Social Movements and the City: Urgent Reflections Amid"Compassion Fatigue", ”Educar Em Revista, vol. 40, 2024. doi:10. 1590/1984-0411. 94770-t.

[4] D. Kranke,B. Kranke,S. Milligan,and A. Dobalian,“Introducing Trauma Trigger Fatigue as an Underlying Factor of Social Work Burnout,”Social Work, vol. 69, no. 4, 2024. doi:10. 1093/sw/swae034.

[5] M. Li,and Q. Ma,“The effect of social media fatigue on compulsive buying,” Behaviour&Information Technology, vol. 44, no. 7, pp. 1429-1445,2025. doi:10. 1080/0144929x. 2024. 2358359.

[6] H. Patzelt,and D. A. Shepherd,“A fatigue model of social venturing,”Small Business Economics,vol. 63, n o. 3, pp. 1065-1088,2024. doi: 10. 1007/s11187-023-00853-4.

[7] W. Ding, Y. Geng, J. Huang, H. Ju, H. Wang, and C. -T. Lin, “MGRW-Transformer:Multigranularity Random Walk Transformer Model for Interpretable Learning,” lEEE Transactions on Neural Networks and Learning Systems,vol. 36, no. 1, pp. 1104-1118, 2025. doi:10. 1109/tnnls. 2023. 3326283.

[8] M. Ekvall,P. Truong,W. Gabriel,M. Wilhelm,and L. Kall, “Prosit Transformer:A transformer for Prediction of MS2 Spectrum Intensities,” Journal of Proteome Research, vol. 21, no. 5, pp. 1359-1364,2022. doi: 10. 1021/acs. jproteome. 1c00870.

[9] Xin Bi and Tian Zhang, "Pedagogical sentiment analysis based on the BERT-CNN-BiGRU-attention model in the context of intercultural communication barriers," PeerJ Computer Science, vol. 10, pp. e2166, 2024. doi:10.7717/peerj-cs.2166.

[10] Tal Yarkoni and Jacob Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning," Perspectives on Psychological Science, vol. 12, no. 6, pp. 1100-1122, 2017.

[11] Nur Hani Zainal, Hui Han Tan, Ryan Yee Shiun Hong, and Michelle Gayle Newman, "Prescriptive predictors of mindfulness ecological momentary intervention for social anxiety disorder: Machine learning analysis of randomized controlled trial data," JMIR Mental Health, vol. 12, no. 1, pp. e67210, 2025.

[12] S. Pradhan,“Social network fatigue: revisiting the antecedents and consequences,” Online Information Review,vol. 46, no. 6, pp. 1115-1131,2022. doi: 10. 1108/oir-10-2020-0474.

[13] J. K. Santiago,M. T. B. Tiago,and F. Tiago,“Social media challenges: fatigue,misinformation and user disengagement,” Management Decision,2025. doi: 10. 1108/md-04-2024-0868.

[14] S. Sunil,M. K. Sharma,S. Amudhan,N. Anand,and N. John,“Social media fatigue:Causes and concerns,” International Journal of Social Psychiatry,vol. 68, no. 3, pp. 686-692,2022. doi:10. 1177/00207640221074800.

[15] T. Wu,and C. -R. Lu,“Understanding compassion fatigue among social workers: a scoping review,” Frontiers in Psychology,vol. 16,2025. doi: 10. 3389/fpsyg. 2025. 1500305.

[16] M. Arboleda-Florez,and C. Castro-Zuluaga,“Interpreting direct sales’ demand forecasts using SHAP values,”Production, vol. 33, pp. e20220035-e20220035, 2023. doi:10. 1590/0103-6513. 20220035.

[17] M. L. Baptista,K. Goebel,and E. M. P. Henriques,“Relation between prognostics predictor evaluation metrics andlocal Interpretability SHAP values,"Artificial Intelligence, vol. 306,2022. doi:10. 1016/j. artint. 2022. 103667.

[18] J. E. Choi,J. W. Shin,and D. W. Shin,“Vector SHAP Values for Machine Learning Time Series Forecasting,” Journal of Forecasting,vol. 44, no. 2, pp. 635-645,2025. doi:10. 1002/for. 3220.

[19] S. Matthews,and B. Hartman, “mSHAP:SHAP Values for Two-Part Models,” Risks,vol. 10, no. 1,2022. doi:10. 3390/risks10010003.

[20] Rahman, M. H., C. Xie & Z. Sha, "Predicting sequential design decisions using the function-behavior-structure design process model and recurrent neural networks," Journal of Mechanical Design, vol. 143, no. 8, pp. 081706, 2021.

[21] Abdesselem Boulkroune, Sarah Hamel, Farouk Zouari, Abdelkrim Boukabou, and Asier Ibeas, "Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities," Mathematical Problems in Engineering, vol. 2017, no. 1, pp. 8045803, 2017. doi: 10.1155/2017/8045803.

[22] Abdesselem Boulkroune, Farouk Zouari, and Amina Boubellouta, "Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems," Journal of Vibration and Control, vol., pp. 10775463251320258, 2025. doi: 10.1177/10775463251320258.

[23] Loubna Merazka, Farouk Zouari, and Abdesselem Boulkroune, "High-gain observer-based adaptive fuzzy control for a class of multivariable nonlinear systems," 2017 6th International Conference on Systems and Control (ICSC), pp. 96-102, 2017. doi: 10.1109/ICoSC.2017.7958728.

[24] G Rigatos, M Abbaszadeh, B Sari, P Siano, G Cuccurullo, and F Zouari, "Nonlinear optimal control for a gas compressor driven by an induction motor," Results in Control and Optimization, vol. 11, pp. 100226, 2023. doi: 10.1016/j.rico.2023.100226.

[25] Farouk Zouari, K Ben Saad, and M Benrejeb, "Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems," International Review on Modelling and Simulations, vol. 5, no. 5, pp. 2075-2103, 2012. https://www.scopus.com/pages/publications/84873265173.

[26] Farouk Zouari, Kamel Ben Saad, and Mohamed Benrejeb, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems", 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), pp. 1-6, 2013. doi: 10.1109/SSD.2013.6564134.

Optimizing Deep Learning Model Ensembles for Plant Disease Detection through Ablation and Correlation Analysis

Nawel Ghrieb[1*], Tahar Guerram[2], Othaila Chergui[3]
[1]Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi university, Tebessa, 12000, Algeria
[2] RelaCS2Laboratory, Larbi Ben M'hidi University, Oum El Bouaghi, 04000, Algeria
[3]Laboratory of Signals and Smart Systems, Echahid Cheikh Larbi Tebessi university, Tebessa, 12000, Algeria
E-mail: nawel.ghrieb@univ-tebessa.dz, tahar_guerram@yahoo.fr, othaila.chergui@univ-tebessa.dz
[*]Corresponding author

Keywords: Plant disease detection, deep learning, ensemble learning, ablation study, vision transformer, Swin transformer, PlantDoc

Received: July 19, 2025

Early detection of plant diseases is crucial for global food security. While Deep Learning ensemble techniques are widely adopted to improve performance, the assumption that simply aggregating models is always beneficial should be nuanced. This paper addresses this issue by conducting a rigorous analysis of an ensemble of four state-of-the-art architectures (Swin Transformer, Vision Transformer, EfficientNetV2, ConvNeXt) on the PlantDoc dataset, a benchmark known for its complexity.
Our approach is twofold. First, we conduct a systematic ablation study to assess how each model contributes to the ensemble's performance. This analysis leads to the counter-intuitive finding that an optimized three-model subset (Swin, ViT, and EfficientNetV2) outperforms the full four-model ensemble. Quantitatively, the pruned ensemble achieves a Macro F1-score of 0.7503 and an accuracy of 0.7619, compared to 0.7409 and 0.7500 for the full set, respectively. Second, to explain this phenomenon, we perform a prediction correlation analysis. This reveals significant predictive redundancy, stemming from the architectural similarities between ConvNeXt and the Transformer-based models, as the cause for this sub-optimality. These findings suggest a key principle for ensemble design: the predictive complementarity of the models is a more critical factor than their individual performance or the complexity of the aggregation method-a finding reinforced by our benchmark showing that even advanced strategies like calibrated voting and stacking failed to outperform the pruned ensemble. Our work thus positions the methodological pairing of ablation study and correlation analysis as an essential and pragmatic approach to optimize the performance of ensembles in computer vision.

Povzetek: Prispevek pokaže, da pri zgodnjem odkrivanju bolezni rastlin ni vedno bolje združiti več modelov: z analizo prispevkov in korelacij napovedi ugotovijo, da manjši, premišljeno izbran ansambel doseže boljše rezultate zaradi manjše redundance.

1 Introduction

Safeguarding crop health is a cornerstone of sustainable agriculture and food security. Plant diseases are responsible for significant yield losses worldwide annually. The rapid and accurate identification of these diseases by experts is often costly, time-consuming, and inaccessible in many regions. Facing this challenge, artificial intelligence, and more specifically Deep Learning, has emerged as a promising solution for automating plant disease diagnosis from leaf images.

From Convolutional Neural Network (CNN) architectures to the more recent Vision Transformers (ViT), numerous models have been successfully applied to this task. To further improve model performance, ensemble techniques, which combine the predictions of multiple models, have become a common practice. The underlying assumption is that models with diverse architectures will learn different feature representations and that their errors will be uncorrelated, thereby allowing the ensemble to correct individual weaknesses.

However, most works merely aggregate high-performing models without deeply analyzing the synergy or conflict between them. The question of whether a larger ensemble is always better remains open. Indeed, the existing literature, while rich in ensemble applications, often focuses on demonstrating overall improvement without systematically quantifying the contribution, whether positive or negative, of each constituent model. Ablation studies, when present, are rarely the main focus of the investigation, leaving the question of optimal ensemble composition largely unexplored. In this context, our contribution is threefold:

1. We evaluate and compare four modern architectures (Swin-Base, ViT-Base, EfficientNetV2-B3, ConvNeXt-Tiny) on the complex data of the PlantDoc dataset.

2. We implement a Soft Voting ensemble method and conduct a rigorous ablation study to quantify the contribution of each component.
3. We demonstrate that careful model selection is paramount and that a smaller, optimized ensemble can outperform a larger one, thereby providing guidelines for designing more effective ensembles.

The remainder of this paper is structured as follows. Section 2 reviews the related work on plant disease detection and the use of ensemble techniques. Section 3 details our systematic methodology, including the data preparation, the advanced training protocol for individual models, and our comprehensive protocol for ensemble optimization which covers the ablation study, the benchmark of alternative aggregation strategies, and the correlation analysis. Section 4 presents the quantitative results of these experiments, detailing the performance of individual architectures and all ensemble configurations, and concludes with an analysis of the optimal model's error patterns. Section 5 then provides an in-depth discussion, contextualizing our results by comparing them to the state of the art, analyzing their broader implications for ensemble design, and outlining the limitations of our study. Finally, Section 6 concludes the paper, summarizing our contributions and suggesting directions for future work.

2 Related work

Plant disease detection has undergone a fundamental transition from manual approaches to automated solutions. Traditional methods, though sometimes still used, are widely recognized as being costly, subjective, and labor-intensive. To overcome these limitations, research has shifted towards techniques that leverage computer vision and image processing, analyzing features such as texture, color, or shape to identify pathologies.

In this context, Machine Learning (ML) offered a first generation of effective tools. While these algorithms were applied to a wide range of agricultural tasks, their use in disease identification specifically highlighted a key limitation. Foundational studies in this area successfully used a variety of classical classifiers. This approach typically relied on a multi-stage process where engineered features like color, texture, and shape were manually extracted from a region of interest before classification. Representative studies include using a Support Vector Machine (SVM) to classify grape leaf diseases based on color and texture features [1], or a K-Nearest Neighbor (KNN) classifier for groundnut leaves after a similar feature extraction stage [2]. Although respectable accuracies were achieved, the performance of all these ML models remained fundamentally dependent on this manual feature engineering step. This process, which requires domain expertise to define relevant features, limited their ability to capture the full complexity of visual patterns and their adaptability to new diseases.

To overcome these obstacles, Deep Learning (DL) has established itself, offering the crucial advantage of automatic feature learning. One of the foundational works in this field was conducted by Mohanty et al., who used an AlexNet-type CNN to achieve 99.27% accuracy on the public PlantVillage dataset, which consists of images taken under controlled laboratory conditions [3]. Too et al. later reinforced these findings in a comparative study, reporting 99.75% accuracy with a DenseNet121 architecture on the same data [4].

This initial success with Convolutional Neural Networks (CNNs) was systematically confirmed by a series of literature reviews, from early analyses [5] up to more recent and comprehensive surveys [6], all establishing the consistent superiority of DL over traditional ML methods. The high effectiveness of this approach is now routinely demonstrated on focused tasks, such as the use of a custom lightweight CNN to detect Bacteriosis in peach leaves with 99% accuracy [7]. The drive for peak performance in such idealized environments has continued with the development of even more sophisticated architectures. A prime example is the work by Zhao et al., [8] who engineered a CNN fusing Inception structure with attention mechanisms to further perfect performance on this standard benchmark.

However, this very success on controlled datasets highlighted the next major challenge for the field: generalizing performance to 'in-the-wild' scenarios. This generalization failure is not trivial; for instance, as noted by Chandra et al., a powerful model like InceptionResNetV2 sees its performance collapse from over 99% on PlantVillage to 39.87% accuracy and a 38% F1-score when evaluated on the real-world complexities of the PlantDoc dataset [9]. To address these limitations, researchers such as Iftikhar et al. [10] have explored the use of optimized CNNs for early detection under challenging real-world conditions. Such scenarios, which involve images with complex backgrounds, variable lighting, and different angles, are precisely the conditions represented by the PlantDoc dataset used in our study.

Despite these emerging solutions, relying on single models remains unreliable due to their high sensitivity to the quality and diversity of training data. The robustness of models against background noise in images, symptom variability, and their effective deployment in the field remain active areas of research. It is precisely to address these challenges of robustness and generalization that research has turned to a complementary strategy: ensemble learning.

Ensemble learning [11], is a well-established strategy for enhancing overall model performance [12]. The core idea is to combine predictions from several models to achieve a more robust and stable final decision. Common techniques range from simple majority voting [13], [14] to more sophisticated strategies like weighting [15], soft voting (averaging probabilities) [16] and stacking [12]. A key limitation of these conventional methods, however, is their primary focus on the aggregation mechanism itself, operating under the implicit assumption that simply combining individually strong models is enough.

This assumption is an oversimplification. The fundamental literature on ensemble learning establishes that the effectiveness of an ensemble depends not so much on the individual performance of its members as on their

predictive diversity [11]. A group of models that make the same mistakes offers no advantage. This principle has led to the development of a vast research field dedicated to ensemble selection (or ensemble pruning), for which the state-of-the-art has been summarized in comprehensive literature reviews such as that of Sagi & Rokach [16]. Within this domain, several families of strategies have been successfully developed and applied. These include, for instance, clustering-based approaches, which were proposed as early as the beginning of the 2000s [17]. The effectiveness of these selection methods has been empirically demonstrated in comparative reviews that evaluate different techniques on complex tasks [18]. Other seminal works have analyzed heuristic search techniques like Backward Elimination, a method whose efficacy was originally demonstrated in a high-stakes domain: the medical diagnosis of pneumonia risk [19].

However, despite the proven utility of these selection methods in other complex domains, their systematic application remains surprisingly rare in the specific context of plant disease detection. The majority of recent studies that do employ ensembles tend to focus on simple aggregation mechanisms. For example, Astani et al. used an ensemble classification approach to achieve high accuracy on tomato diseases [20], and Shafik et al. introduced ensemble techniques to enhance model robustness [21]. While these works demonstrate the benefits of combining models, they operate under the implicit assumption that aggregating individually strong models is sufficient. This approach overlooks the crucial question of predictive redundancy and fails to investigate the internal composition of the ensemble itself.

This need for complementarity is particularly relevant with the emergence of state-of-the-art architectures like the Vision Transformer (ViT) [22], the Swin Transformer [23], and modern CNNs like ConvNeXt [24]. These models are rapidly being adopted for disease classification, with much of the research focusing on maximizing their individual performance. This trend is evident across different architectural families. In the domain of modern CNNs, for instance, the study by Chen et al. demonstrated that integrating attention mechanisms into a MobileNetV2 backbone could significantly improve accuracy for rice disease identification [25]. Similarly, for Vision Transformers, research has focused on exhaustive hyperparameter tuning, as illustrated by the study by Ouamane et al., which achieved 99.77% accuracy on the controlled PlantVillage dataset through an optimized ViT configuration [26]. Building on this, some studies have explored combining these architectures into fixed hybrid systems. Notable examples include the "PlantXViT" introduced by Thakur et al. [27] and the "TLMViT" model from Tabbakh et al. [28], both of which created monolithic models by fusing Transformer and CNN-based components.

However, while these studies showcase the power of these new architectures, they concentrate on their performance either in isolation or as fixed hybrid systems. How these models interact within a multi-model ensemble, and how to determine its optimal composition, remains a largely unexplored area. The fact that an optimized ConvNeXt is powerful does not explain how its predictions correlate with those of a Swin Transformer or a ViT within the same team. Understanding whether these architectures are predictively redundant or complementary is a crucial question. To our knowledge, no study has conducted a systematic ablation analysis coupled with a correlation study to empirically investigate this redundancy and complementarity among these specific architectural families for plant disease detection under complex conditions.

Our contribution aims precisely to fill this gap by providing empirical evidence that a carefully pruned, heterogeneous ensemble (Transformer + CNN) can outperform the full set of models by eliminating predictive redundancy. While the methods we employ (ablation and correlation) are conventional, their application to this specific problem yields novel insights into the optimal composition of modern vision model ensembles, providing pragmatic guidelines for practitioners.

3 Proposed methodology

Our methodological approach is designed first to maximize the performance of each individual model, and then to systematically optimize the ensemble's composition. The core pipeline for this process, illustrated in Figure 1, is broken down into several stages: data preparation (3.1, 3.2), a rigorous training protocol for four state-of-the-art architectures (3.3), and the construction of an ensemble optimized via a systematic ablation study (3.4 and 3.5). This empirical optimization is then explained through a final correlation analysis, also detailed in section 3.5.

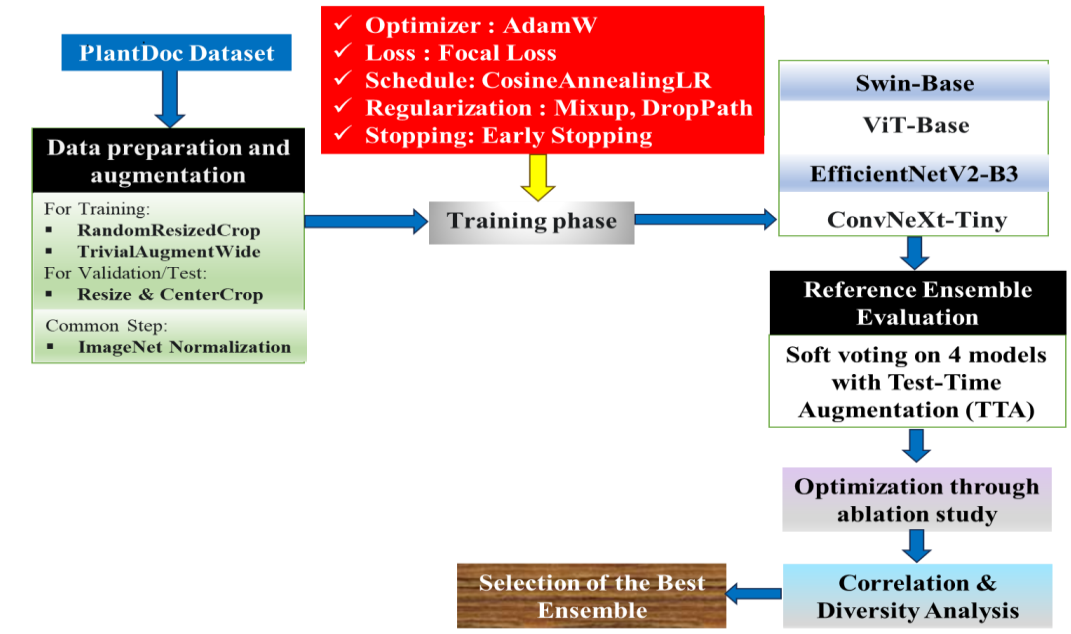


Figure 1: Methodological pipeline for ensemble construction and optimization

3.1 Dataset

Our study is conducted on the public PlantDoc dataset [29], which contains leaf images of 13 different plant species, divided into 27 classes that include both healthy leaves and various disease classifications—see Figure 2 for the distribution and composition of the dataset.

The majority of the images in the PlantDoc dataset were captured in the field (in-the-wild) using smartphones. Unlike other datasets that use simple backgrounds, these images were taken in a natural environment with complex backgrounds, under variable lighting conditions, and from different angles [29], which presents a more realistic challenge for detection models. Image samples illustrate this diversity—see also Figure 3 for image examples.

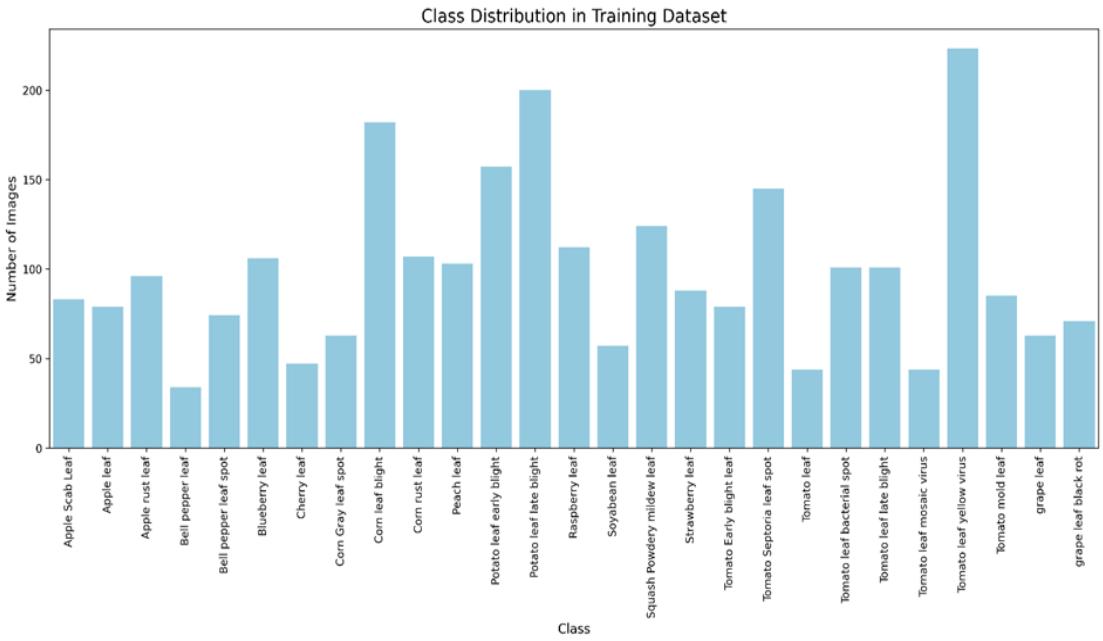
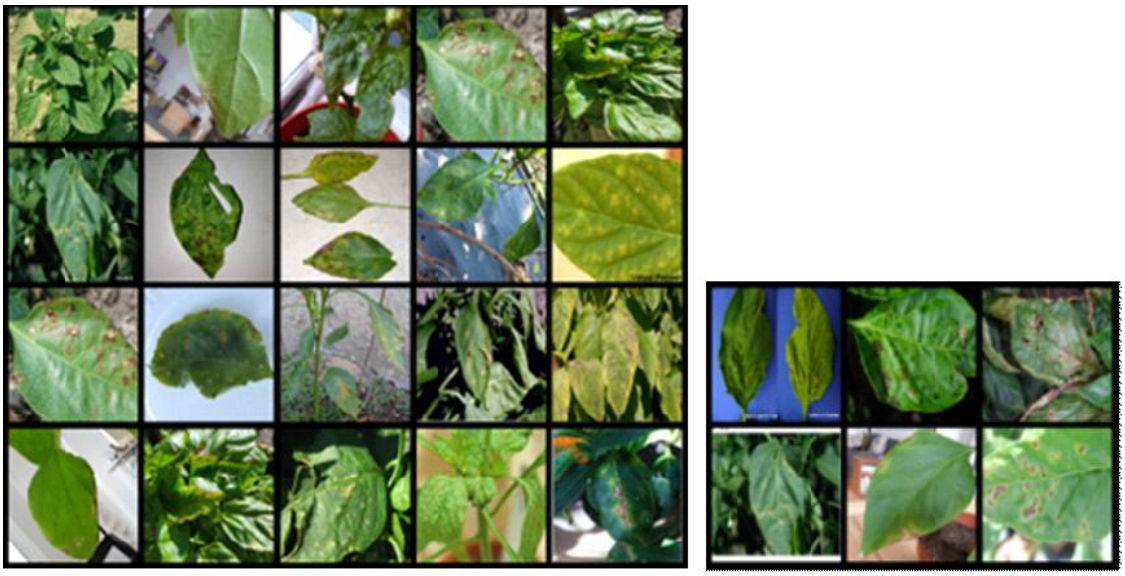


Figure 2: Class distribution in the PlantDoc training set



(a) Train subset of Bell pepper leaf spot class.

(b) Test subset of Bell pepper leaf spot class.

Figure 3: Sample images of PlantDoc Bell pepper leaf spot class.

The dataset is pre-divided into a training set and a test set. In accordance with the official split, we use the test set for our final evaluation. Upon verification of the publicly available data archive, we identified 252 usable image files, which we used for all our experiments to ensure reproducibility. For the training of each individual model, we subdivided the provided training set, allocating 85% of the images to a new training set and the remaining 15% to a validation set. To ensure a balanced representation of classes, this division was performed using stratification. The exact per-class counts for each split are detailed in Appendix A (Table A1). To ensure full auditability, the complete list of filenames used in the test set is provided in Appendix B (Table B1).

3.2 Data preprocessing and augmentation

Our data pipeline distinguishes between the training and evaluation phases to maximize robustness and ensure consistent testing.

For training, an aggressive data augmentation strategy was employed to improve the models' robustness. The transformations applied to the training images include a RandomResizedCrop to 224x224 pixels and the application of TrivialAugmentWide, an automatic augmentation policy that combines multiple operations.

For the validation and test sets, deterministic preprocessing was used to ensure consistent evaluation. The images were first resized so that their shorter side was 224 pixels while maintaining the aspect ratio, and then a 224x224 CenterCrop was applied. All images, after transformation, were converted to PyTorch tensors and normalized using the mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225], which are standard for models pre-trained on ImageNet.

Furthermore, to ensure a robust evaluation and mitigate sensitivity to object centering, all final performance metrics on the test set were obtained using Test-Time Augmentation (TTA). For each test image, we averaged the predictions from the original image and its horizontally flipped version.

3.3 Architectures and individual model training

To ensure a rich and informative comparison, we selected four state-of-the-art models, summarized in Table1, chosen specifically to represent diverse architectural families. This selection includes:

- Two prominent Transformer-based architectures: The Swin Transformer (Swin-Base) and the standard Vision Transformer (ViT-Base).
- Two modern CNN-based architectures: A recent version of EfficientNet (EfficientNetV2-B3), known for its improved training efficiency and performance, and a modern CNN (ConvNeXt-Tiny) that incorporates architectural insights from Transformers.

This strategic selection enables a rigorous analysis of predictive synergies and conflicts both within and between architectural families. To establish a fair and equitable comparison, all models were initialized with weights pre-trained on the large-scale ImageNet-21k dataset (or the comparable ImageNet-22k). The exact timm model identifiers used for each architecture are detailed in Table 1 to ensure full reproducibility.

Table 1: Model architectures used in the study

Model	Timm Identifier	Type	Description
Swin-Base	swin_base_patch4_window7_224.ms_in22k	Transformer	Hierarchical Transformer using shifted attention windows.
ViT-Base	vit_base_patch16_224.augreg_in21k	Transformer	Standard Vision Transformer architecture.
EfficientNetV2-B3	tf_efficientnetv2_b3.in21k	CNN	Modern convolutional network with fused-MBConv blocks.
ConvNeXt-Tiny	Convnext tiny.fb_in22k	CNN	Modern convolutional network inspired by Transformers.

Subsequently, each model was fine-tuned independently on the PlantDoc training set following a rigorous and unified training protocol. To ensure the robustness of our findings and to account for the variance inherent in the training process, each individual model training was conducted three times using distinct random seeds (42, 50, and 72). The mean and standard deviation of these runs are reported in Section 4.1 to demonstrate the stability of our models. All subsequent analyses in Section 4 (including the ensemble studies and benchmarks) are then performed on the single, most representative set of models (from seed 42) to ensure clarity and consistency. This multi-seed validation approach ensures that our main conclusions are drawn from a stable and representative baseline, rather than from a single, potentially fortunate outcome.

The training protocol was designed to maximize the performance of each model before their integration into the ensemble, and its key components are as follows:

- **Optimizer:** We used the AdamW optimizer with an initial learning rate of 5e-5 and a weight decay of 0.05.
- **Loss function**: To effectively address the potential class imbalance in the PlantDoc dataset, we opted for the Focal Loss with parameters alpha=0.25 and gamma=2.0
- **Learning rate scheduler:** A CosineAnnealingLR scheduler was implemented, preceded by a linear warmup phase over the first 5 epochs to stabilize initial training.
- **Regularization:** In addition to data augmentation, we employed the Mixup technique (alpha=0.5) during training to improve the model's generalization by creating virtual examples through linear interpolation of images and their labels. DropPath regularization (rate=0.2) was also enabled for architectures that support it.
- **Stopping conditions and model selection**: Training was conducted for a maximum of 60 epochs. An Early Stopping mechanism was implemented to prevent overfitting: if the validation loss did not improve for 10 consecutive epochs, training was stopped. For the final evaluation and inclusion in the ensemble, we kept the model weights from the epoch that achieved the lowest validation loss. This ensures that each model is evaluated at its peak performance.

The chosen hyperparameters are based on established best practices for fine-tuning ImageNet-pretrained Transformer-based and modern CNN models. The learning rate of 5e-5 is a common and effective starting point for the AdamW optimizer, while the Focal Loss parameters (alpha=0.25, gamma=2.0) correspond to the widely adopted default values established by Lin et al. [30], known for their robustness in handling class imbalance.

This rigorous and unified training protocol was systematically applied to each of the four base architectures. The use of an Early Stopping mechanism based on validation performance ensures that every model was trained to its peak performance under identical conditions, guaranteeing a fair comparison under identical optimization constraints. Consequently, the contribution of each model in the subsequent ensemble study is a function of its intrinsic capabilities and complementarity, rather than its training duration.

3.4 Ensemble method: soft voting

To aggregate the predictions from our models, we employ Soft Voting, a classic and effective ensemble method [16]. The process consists of two main steps.

First, for a given input image, each individual model I within the ensemble E produces a vector of raw prediction scores, known as logits (z_i). This logit vector is then converted into a probability vector p_i using the Softmax function, ensuring that the outputs represent a valid probability distribution over the C classes. The probability for the class j from the model i is calculated as follows:

$$p_{i,j} = \frac{e^{z_{i,j}}}{\sum_{k=1}^{C} e^{z_{i,k}}} \quad (1)$$

Where:

$\boldsymbol{p_{i,j}}$ is the probability assigned by model i to class j.

$\boldsymbol{z_{i,j}}$ is the logit (raw score) from model i for class j.

C is the total number of classes.

k is an index used to iterate over all classes for the normalization sum.

Second, the final probability vector for the ensemble, P_{ens}, is obtained by computing the element-wise average of the individual probability vectors from all models in the ensemble:

$$P_{ens} = \frac{1}{|E|} \Sigma_{i \in E} p_i \quad (2)$$

Where:
P_{ens} is the final averaged probability vector from the ensemble.
E is the set of all models in the ensemble.
|E| is the total number of models in the ensemble.
p_i is the probability vector from model i.

Finally, the class predicted by the ensemble, ŷ, is determined by identifying the class with the highest averaged probability. This is achieved using the argmax function:

$$\hat{y} = \text{argmax_j } (P_{ens,\, j}) \quad (3)$$

Where:
ŷ is the final predicted class label.
argmax_j returns the class index j that maximizes the value of the ensemble probability for that class.

3.5 Ensemble optimization protocol

To systematically identify the optimal ensemble composition, we implemented a multi-stage analysis protocol. Our approach adapts the principle of ablation studies—a standard method for assessing component importance—to evaluate the marginal contribution of each model within the ensemble [31]. The protocol proceeds as follows:

1. **Full ensemble reference:** First, we construct a reference ensemble (E_full) comprising the four fully optimized trained models (Swin-Base, ViT-Base, EfficientNetV2-B3, ConvNeXt-Tiny). Predictions are aggregated using Soft Voting (Section 3.4), and its performance on the test set is established as our benchmark for the ablation study.
2. **Systematic "Leave-One-Out" ablation**: Building on this reference ensemble, we then perform a systematic ablation. For each model m in the full ensemble, we create a subset (E_full \ {m}) and evaluate its performance. This process allows us to empirically identify the best-performing subset and to quantify the marginal contribution of each model.
3. **Correlation analysis for a deeper explanation**: To understand and explain the results of the ablation study, we quantify the predictive redundancy between the optimized base models. We calculate a Pearson correlation matrix based on the final class predictions generated by each model pair. We intentionally perform this analysis on the official test set. Since the performance of the ablated ensembles is evaluated on this same set, this choice creates a direct and interpretable link between the observed performance changes (the effect) and the predictive behavior of the models on the exact data where those changes occurred (the cause). This provides a clear and methodologically sound explanation for the observed synergy or redundancy.

3.6 Evaluation of alternative aggregation strategies

While our primary optimization method focuses on selecting the best architectural composition of the ensemble (as detailed in Section 3.5), a common alternative is to optimize the aggregation mechanism itself. To provide a robust benchmark, we implemented and evaluated two distinct families of aggregation-focused strategies on the full 4-model ensemble: Soft Voting and Stacking.

3.6.1 Soft voting strategies

As modern deep learning models are frequently miscalibrated (i.e., overconfident), our analysis of soft voting goes beyond a simple average. We systematically evaluated four strategies to assess the impact of probability calibration and advanced weighting schemes. This experiment, conducted on the full four-model ensemble (from seed 42), evaluates and compares the following strategies:

- **Uncalibrated soft voting**: This strategy corresponds strictly to the "Full Ensemble Reference" (4 models) established in our ablation study (Section 3.5). It computes the simple average of the raw probability outputs from each individual optimized model. Its performance serves as the reference point for this comparative analysis.
- **Calibrated soft voting:** In this strategy, each model's probabilities are first calibrated using Temperature Scaling on the validation set. The final prediction is the simple average of these calibrated probabilities.
- **Calibrated Weighted (F1):** This approach applies a weighted average to the calibrated probabilities. The weights are derived from each model's macro F1-score on the validation set.
- **Calibrated Weighted (Logloss):** This strategy also uses a weighted average of calibrated probabilities. To ensure numerical stability and to regularize the weights, they are derived from the inverse of each model's Log-Loss on the validation set, with a small regularization constant ε (epsilon) added to the denominator (weight = 1 / (log_loss + ε)).

3.6.2 Stacking

A more complex, two-level learning approach is implemented. A meta-model (XGBoost) is trained to learn how to best combine the predictions generated by the four optimized base models. To prevent data leakage and ensure a robust evaluation, we employed a rigorous protocol defined by four key components:

1. **Feature generation:** The training features for the meta-learner were generated using a 5-fold stratified cross-validation strategy on the full training set. This created "out-of-fold" predictions, ensuring that the features used to train the meta-learner were generated by base models that had not seen that data during their training within that fold. This strict separation of training and validation flows effectively addresses the overfitting issues observed in simpler stacking implementations.
2. **Input features:** The features provided to the meta-learner were the full probability vectors (outputs of the softmax function) generated by the models trained within the cross-validation loop described in the Feature Generation stage. This resulted in a feature vector of size 108 (4 models × 27 classes) for each training sample.
3. **Base models configuration:** The base models integrated into this stacking framework utilize the optimized hyperparameter configurations established in Section 3.3. To prevent data leakage, a strict distinction was applied regarding model training:
   - **For meta-learner training (Step 1):** While the hyperparameters (e.g., learning rate, batch size, loss function) remained fixed to the optimal values determined in Section 3.3, fresh instances of the base models were re-trained from scratch within each fold of the cross-validation process described in Step 1. This ensures that the features fed to the meta-learner were unbiased (Out-Of-Fold).
   - **For final test inference:** Once the meta-learner was fully tuned, we utilized the distinct, optimized models (corresponding to the representative Seed 42 run detailed in Section 4.1) to generate the final predictions on the test set.
4. **Hyperparameter tuning and final training:** Finally, the hyperparameters of the XGBoost model were themselves optimized using a nested, inner-loop 3-fold cross-validation with GridSearchCV. The optimal configuration identified during this search was then fitted on the complete set of "out-of-fold" predictions to create the final meta-learner.

3.7 Evaluation metrics

To provide a comprehensive evaluation of our models, we used a set of standard classification metrics [32] for multi-class problems. We report three key metrics: Accuracy, Balanced Accuracy (BA), and the Macro F1-score. These metrics are derived from the standard components of a confusion matrix: True Positives (TP), False Positives (FP), and False Negatives (FN).

Accuracy: This metric measures the overall fraction of correctly classified instances and is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

In our multi-class context, per-class metrics are calculated by treating each class c individually (the "one-vs-rest" approach). From these per-class scores, we derive the following aggregated metrics:

Balanced Accuracy (BA): This metric is the average of recall obtained on each class. It is particularly suited for imbalanced datasets as it gives equal weight to each class, regardless of its number of samples. It is defined as:

$$BA = \frac{1}{N} \Sigma_{i=1}^{N} \text{Recall}_i \quad (5)$$

Macro F1-Score: This metric is the unweighted average of the F1-scores calculated for each individual class. By treating all classes equally, it provides a robust measure of a model's ability to perform well across both majority and minority classes. The F1-score for a single class is the harmonic mean of Precision and Recall:

$$\text{F1-Score}_i = 2.\frac{\text{Precision}_i \,.\, \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (6)$$

The Macro F1-Score is then calculated as:

$$\text{Macro F1-Score} = \frac{1}{N} \Sigma_{i=1}^{N} \text{F1-Score}_i \quad (7)$$

where for each class i, N is the total number of classes, and:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (8)$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (9)$$

To ensure a thorough and transparent evaluation, we leverage all three metrics. We use Accuracy for a clear measure of overall correctness and for comparison with the state of the art, while the Macro F1-score and Balanced Accuracy provide crucial insights into the model's fairness and performance on rare classes. This comprehensive approach best reflects our goal of developing a robust and effective model.

3.8 Experimental setup

The experiments were conducted on Python 3.10 using the PyTorch framework (v2.5.1) and the timm library (v1.0.16). The evaluation was performed on an NVIDIA RTX 4060Ti GPU with a batch size of 16, utilizing CUDA v12.1. Performance metrics were calculated using the scikit-learn library (v1.7.2) on the official PlantDoc test set. The reproducibility of our results was ensured by setting a fixed random seed for all relevant libraries in each experiment. The specific seeds used are detailed in the corresponding results sections (e.g., Section 4.1).

4 Results

This section presents the results of our experiments, structured to systematically identify and analyze the optimal model ensemble. First, we establish the individual performance of each state-of-the-art architecture (Section 4.1). Second, we identify the best-performing ensemble configuration through a two-part analysis (Section 4.2): a "leave-one-out" ablation study to find the optimal model subset, followed by a benchmark against alternative aggregation strategies. Third, to explain the ablation results, we conduct an analysis of synergy and redundancy using correlation and diversity metrics (Section 4.3). Finally, we perform an in-depth, multi-level error analysis of the best-performing ensemble to understand its strengths and weaknesses (Section 4.4).

Across all stages of this evaluation, performance is measured using a comprehensive set of metrics-Accuracy, Balanced Accuracy, and Macro F1-score-whose complementary roles and definitions are detailed in Section 3.7.

4.1 Individual model performance

The evaluation of individual model performance is the first step in our analysis. To rigorously assess performance and stability, and to account for the variance inherent in the training process, each architecture in our optimized protocol (Section 3.3) was trained and evaluated three times using distinct random seeds (42, 50, and 72). The detailed per-seed results are presented in Table 2. To synthesize these findings and facilitate a clearer comparison of overall performance and stability, the mean and standard deviation for each metric were calculated and are presented in Table 3.

Analysis of the multi-seed results reveals a clear performance hierarchy. The summary in Table 3 shows that the Swin Transformer (Swin-Base) emerges as the top-performing model across all average metrics, achieving a mean accuracy of 73.81% and the highest mean F1-score (macro) of 72.72%. The ConvNeXt-Tiny follows as a very strong competitor. In contrast, while the Vision Transformer (ViT-Base) shows the highest stability with the lowest standard deviations (e.g., ±0.0040 for accuracy), its average performance is lower than the other two main architectures. As detailed in Table 2, the absolute peak F1-score of the study (0.7418) was achieved by ConvNeXt-Tiny with seed 50, while the peak accuracy (0.7460) was achieved by Swin-Base with seed 42.

For the subsequent in-depth analyses, including the ensemble study, it was necessary to select a single, representative set of models from our multi-seed runs. Our choice fell on the models generated with seed 42, as a rigorous analysis showed this seed to be the most representative of the models' typical behavior (i.e., its results had the lowest average distance to the mean performance across all metrics). While this choice was driven by objectivity, it is worth noting that this seed also produced the study's single highest accuracy score.

Table 2: Detailed performance per model and random seed.

Model	Seed	Accuracy	Balanced Accuracy	Macro F1-score
Swin-Base	42	0.7460	0.7434	0.7412
	50	0.7341	0.7350	0.7293
	72	0.7341	0.7242	0.7110
ViT-Base	42	0.7063	0.7073	0.7009
	50	0.6984	0.6959	0.6844
	72	0.7024	0.7036	0.6909
EfficientNetV2-B3	42	0.6032	0.5913	0.5854
	50	0.5635	0.5691	0.5488
	72	0.6151	0.6160	0.6007
ConvNeXt-Tiny	42	0.7183	0.7102	0.7062
	50	0.7302	0.7389	0.7418
	72	0.7103	0.7069	0.7014

Table 3: Mean performance and stability of optimized models (over 3 seeds).

Model	Accuracy (Mean ± Std. Dev.)	Balanced Accuracy (Mean ± Std. Dev.)	Macro F1-score (Mean ± Std. Dev.)
Swin-Base	0.7381 ± 0.0069	0.7342 ± 0.0096	0.7272 ± 0.0152
ViT-Base	0.7024 ± 0.0040	0.7023 ± 0.0058	0.6921 ± 0.0084
EfficientNetV2-B3	0.5939 ± 0.0270	0.5922 ± 0.0235	0.5783 ± 0.0267
ConvNeXt-Tiny	0.7196 ± 0.0100	0.7186 ± 0.0176	0.7165 ± 0.0221

Having established the robustness of our models and selected a representative configuration, we then quantified the effectiveness of our training protocol (described in Section 3.3). To do so, Table 4 compares the performance of the "Optimized" models against a "Baseline" configuration. To ensure a fair and direct comparison, both configurations were run using the same representative random seed (seed 42). The Baseline

configuration uses a minimal setup (e.g., Cross-Entropy loss, fixed learning rate).

Table 4: Performance gain of the optimized protocol (seed 42) vs. baseline.

Model	Accuracy	Balanced Accuracy	Macro F1-score
Swin-Base (Baseline)	0.7222	0.7325	0.7182
Swin-Base (Optimized)	0.7460 (+3.3%)	0.7434 (+1.5%)	0.7412 (+3.2%)
ViT-Base (Baseline)	0.6389	0.6403	0.6342
ViT-Base (Optimized)	0.7063 (+10.5%)	0.7073 (+10.5%)	0.7009 (+10.5%)
ConvNeXt-Tiny (Baseline)	0.7063	0.7053	0.6933
ConvNeXt-Tiny (Optimized)	0.7183 (+1.7%)	0.7102 (+0.7%)	0.7062 (+1.9%)
EfficientNetV2-B3 (Baseline)	0.5714	0.5820	0.5653
EfficientNetV2-B3 (Optimized)	0.6032 (+5.6%)	0.5913 (+1.6%)	0.5854 (+3.6%)

Analysis of Table 4 reveals two major and complementary conclusions.

First, the application of our optimized protocol leads to a systematic and significant performance improvement. The gain on the macro F1-score is particularly notable, with an increase of +10.5% for ViT-Base and +3.2% for the top-performing Swin-Base. This result convincingly validates the effectiveness of our approach, which combines a Focal Loss function to address class imbalance, an advanced learning rate scheduler (CosineAnnealingLR with warmup), and aggressive regularization techniques (Mixup, TrivialAugmentWide, and DropPath) that effectively combat overfitting and improve generalization.

Second, the comparison highlights a clear and consistent performance hierarchy among the architectures. The Swin Transformer establishes itself as the top-performing model under both protocols, setting a strong benchmark. Conversely, EfficientNetV2-B3 consistently displays the most modest performance. This performance hierarchy, observed under identical protocol conditions, suggests that modern architectures capable of modeling global context or incorporating Transformer-like design principles (such as Swin and ConvNeXt) are intrinsically well-suited to the complexities of the PlantDoc dataset, compared to standard CNNs. These optimized individual performances will serve as a solid foundation for the next step: our ensemble construction and ablation analysis.

4.2 Ensemble optimization: composition vs. aggregation strategy

To identify the optimal ensemble configuration, we conducted a two-part comparative analysis. First, we performed a "leave-one-out" ablation study to find the best subset of models by optimizing the ensemble's composition. Second, we evaluated alternative strategies that optimize the aggregation mechanism itself as a benchmark.

4.2.1 Optimizing ensemble composition via ablation study

The results of our "leave-one-out" ablation study, presented in Table 5, demonstrate a key principle of ensemble learning: adding more models does not always improve performance. On the contrary, our analysis reveals that a three-model subset performs better than the full ensemble.

Table 5: Performance of the ablation study ensembles (soft voting)

Ensemble Configuration	Accuracy	Balanced Accuracy	Macro F1-score
ABLATION STUDY (Soft Voting)			
Ensemble (full, 4 models)	0.7500	0.7433	0.7409
Ensemble (without ViT-Base)	0.7579	0.7543	0.7500
Ensemble (without EfficientNetV2-B3)	0.7540	0.7464	0.7411
Ensemble (without ConvNeXt-Tiny)	**0.7619**	**0.7542**	**0.7503**
Ensemble (without Swin-Base)	0.7222	0.7150	0.7076

The primary finding is that the ensemble without the ConvNeXt-Tiny model achieved the highest scores across all primary metrics, with an accuracy of 0.7619 and a macro F1-score of 0.7503. This configuration outperforms the full four-model ensemble, which peaks at an accuracy of 0.7500 and an F1-score of 0.7409. This improvement suggests that the predictions from the ConvNeXt-Tiny model, while strong individually, introduced a degree of redundancy or predictive noise that limited the collective performance. This redundancy hypothesis will be further investigated in Section 4.3 through the analysis of the prediction correlation matrix.

This ablation study also allowed us to quantify the importance of each architecture. The critical contribution of the Swin-Base is highlighted by the drastic performance drop (F1-score to 0.7076) observed upon its removal, identifying it as the cornerstone of our ensemble.

While the superiority of the three-model ensemble is numerically clear, a formal confirmation of its statistical significance would require a larger test set. Nevertheless, these results provide strong empirical evidence that the careful selection of a diverse and complementary subset of models is a more effective optimization strategy than simply aggregating all available models. For a detailed, per-class performance breakdown of the best-performing ensemble (without ConvNeXt-Tiny) and the full four-model ensemble, refer to the full classification reports in Appendix A (Table A2 and Table A3, respectively).

4.2.2 Benchmarking against alternative aggregation strategies

To further validate our selection-based optimization approach, we benchmarked the performance of our best ablation ensemble (Ensemble without ConvNeXt-Tiny) against advanced aggregation strategies applied to the full four-model ensemble. The goal was to determine whether optimizing the aggregation mechanism could match or exceed the performance of optimizing the ensemble's composition. Two families of strategies were tested: advanced soft voting methods (including calibration and weighting) and a two-level Stacking ensemble with an optimized XGBoost meta-learner. The aggregated results are summarized in Table 6, while the detailed, per-class classification reports for each of these benchmarked strategies can be found in Appendix A (Tables A4-A7).

Table 6: Performance benchmark of alternative aggregation strategies on the full ensemble.

Ensemble Strategy	Accuracy	Balanced Accuracy	Macro F1-score
Ensemble (without ConvNeXt-Tiny)	0.7619	0.7542	0.7503
AGGREGATION STRATEGIES (Full Ensemble)			
Uncalibrated Soft Voting	0.7500	0.7433	0.7409
Calibrated Soft Voting	0.7460	0.7402	0.7387
Calibrated Weighted Voting (F1)	0.7500	0.7433	0.7418
Calibrated Weighted Voting (LogLoss)	0.7540	0.7476	0.7481
Stacking (Optimized XGBoost)	0.7222	0.7148	0.7201

As shown in Table 6, while the probability calibration and weighted voting strategies yielded results comparable to the Full Ensemble Reference (Uncalibrated Soft Voting), the complex Stacking approach (implemented via the rigorous leakage-safe protocol described in Section 3.6.2) actually led to a performance degradation (Accuracy: 0.7222, Macro F1: 0.7201). Our pruned three-model ensemble (Ensemble without ConvNeXt-Tiny) achieves the highest scores across all primary metrics, with an accuracy of 0.7619 and a macro F1-score of 0.7503.

This finding is significant. It suggests that when base models exhibit a degree of predictive redundancy (as will be demonstrated in Section 4.3 with the correlation matrix), the performance gain from physically removing a less synergistic model (like ConvNeXt-Tiny) is greater than the gain from applying a more sophisticated weighting or learning scheme that attempts to mitigate its influence. This limitation aligns with the theoretical understanding of meta-learning: stacking algorithms typically require significantly larger validation sets to effectively model complex inter-classifier correlations without overfitting. In scenarios with moderate dataset sizes (as is common in plant disease classification), simpler voting strategies often prove more resilient and generalize better than high-variance meta-learners.

Ultimately, this benchmark demonstrates that the path to an optimal ensemble in our case lies not in more complex aggregation logic, but in the strategic curation of its components. The fact that a smaller, carefully selected ensemble outperforms both a larger one and more complex, advanced aggregation schemes reinforces our fundamental finding.

4.3 Analysis of ensemble synergy and redundancy

The superior performance of the ensemble after removing the ConvNeXt-Tiny model (as shown in Section 4.2) suggests a phenomenon of predictive redundancy. To investigate this, we analyzed the synergy between the optimized base models. This analysis is twofold: a visual assessment via a prediction correlation matrix and a quantitative evaluation using pairwise diversity metrics.

First, we calculated the Pearson correlation between the final class predictions of each of the four architectures on the test set. The resulting correlation matrix is presented in Figure 4.

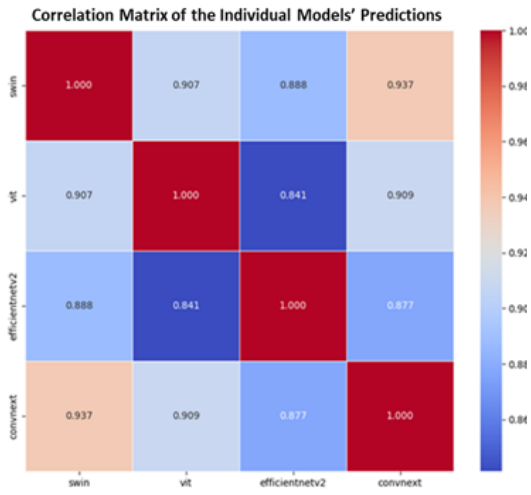


Figure 4: Pearson correlation matrix of model predictions. high values (warm colors) indicate that two models tend to make similar predictions.

The matrix provides crucial insights into the models' predictive behaviors. We observe a very high correlation of 0.937 between Swin-Base and ConvNeXt-Tiny. This strong similarity is likely rooted in their architectural design; ConvNeXt-Tiny was explicitly designed by adapting modern CNNs to incorporate principles from Vision Transformers like Swin-Base, making it plausible that both models learn similar hierarchical feature representations. Conversely, EfficientNetV2-B3 and ViT-Base exhibit the lowest correlation (0.841), highlighting them as the most diverse pair.

To quantify these relationships more precisely, we computed pairwise diversity metrics, including the Disagreement Measure and the Double-Fault Measure. The Disagreement score represents the percentage of test samples on which two models differ, while the Double-Fault score measures the percentage of samples that both models misclassify. The results are summarized in Table7.

Table 7: Pairwise diversity metrics.

Model Pair	Disagreement	Double-Fault
Swin-Base vs Vit-Base	0.1905	0.1786
Swin-Base vs EfficientNetV2-B3	0.1984	0.2262
Swin-Base vs ConvNeXt-Tiny	0.1468	0.1944
Vit-Base vs EfficientNetV2-B3	0.2063	0.2421
Vit-Base vs ConvNeXt-Tiny	0.1786	0.1984
EfficientNetV2-B3 vs ConvNeXt-Tiny	0.2262	0.2262

These metrics, when considered alongside the individual performance results (Table 4), provide a clear justification for our ablation study findings. They reveal a classic ensemble trade-off.

First, the Swin-Base and ConvNeXt-Tiny pair, despite being the two top-performing architectures individually across all evaluated metrics., exhibit the lowest Disagreement score (0.1468). This confirms that despite being individually the strongest models, their predictive similarity limits the marginal benefit of including both in a simple averaging ensemble. Crucially, the second lowest disagreement is observed between ViT-Base and ConvNeXt-Tiny (0.1786). This pattern indicates a systemic redundancy: ConvNeXt-Tiny strongly correlates with both Transformer-based models. Furthermore, it is notable that the disagreement between the two Transformers themselves (Swin vs. ViT, 0.1905) is higher than the disagreement between ConvNeXt-Tiny and either Transformer. This confirms that ConvNeXt-Tiny acts as a redundant architectural intermediary, whereas the Swin-ViT pair preserves greater predictive diversity.

Second, the Double-Fault metric clarifies the synergy within the retained models and further validates the exclusion of ConvNeXt-Tiny. Notably, the Swin-Base vs. ViT-Base pair achieves the lowest Double-Fault score (0.1786) across the entire matrix. This is lower than the Swin-ConvNeXt pair (0.1944), indicating that Swin-Base and ViT-Base are less likely to err simultaneously than Swin-Base and ConvNeXt-Tiny. This makes the Swin-ViT pair the most robust core for the ensemble.

Regarding EfficientNetV2-B3, while its pairs show higher double-fault rates (reflecting its lower individual accuracy on difficult samples), this is counterbalanced by its high disagreement scores with the retained models (e.g., 0.2063 against ViT-Base). This confirms its role as a specialist model that provides diversity, whereas the Swin-ViT pair ensures stability

This detailed analysis explains why removing ConvNeXt-Tiny was the most effective optimization. The optimal ensemble composition (Swin-Base, ViT-Base, EfficientNetV2-B3) is successful because it retains the strongest model (Swin-Base), pairs it with the model least likely to share its errors (ViT-Base), and adds the most diverse model (EfficientNetV2-B3) for a unique corrective perspective. This process eliminated the model most redundant with our core performer, thereby maximizing the overall diversity and error-correction capability of the final ensemble.

Ultimately, this analysis confirms that a principled approach is essential for optimizing ensembles. In our case, the path to the optimal ensemble involved pruning the model (ConvNeXt-Tiny) that offered the least diversity relative to our strongest model (Swin-Base), thereby maximizing the trade-off between high individual performance and collective predictive synergy.

4.4 In-depth analysis of the optimal model's behavior

To understand the performance of our best-performing model, the Ensemble (without Convnext-Tiny), we conducted a multi-level error analysis. This involved a quantitative assessment using the confusion matrix and per-species metrics, as well as a qualitative analysis of failure cases using Grad-CAM to validate our hypotheses.

4.4.1 Quantitative error analysis

The confusion matrix for our best-performing model (Ensemble (without Convnext)) is presented in Figure 5. It enables a finer-grained analysis of the classification errors.

Analysis of the matrix reveals several key points:

1. Well-identified classes: The model demonstrates excellent performance for a significant number of classes. It achieves a perfect recall of 100% for 6 out of 27 classes: Corn_rust_leaf, Raspberry_leaf, Squash_Powdery_mildew_leaf, Strawberry_leaf, grape_leaf and grape_leaf_black_rot. Furthermore, it shows strong discriminative power for other important classes such as Tomato_leaf_yellow_virus (93.3% recall), Tomato_Septoria_leaf_spot (91.7% recall), Apple_Scab_Leaf (90.0% recall), and Apple_rust_leaf (90.0% recall), indicating that their visual features are distinct and well-learned by the ensemble.

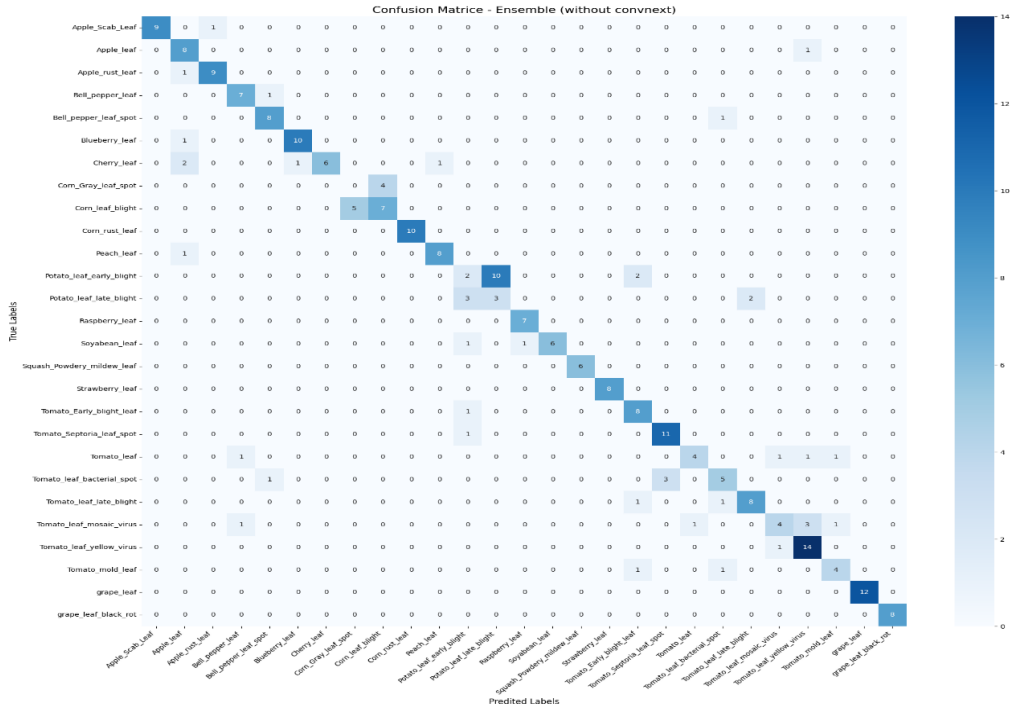


Figure 5: Confusion matrix for the optimal ensemble (without convnext).

2. Main areas of confusion: The matrix highlights significant areas of confusion, which are primarily concentrated within species that have multiple, visually similar diseases.

- **Critical failure on corn_gray_leaf_spot**: The most striking finding is the model's complete failure to identify Corn_Gray_leaf_spot, which has a recall and F1-score of 0.0000. This indicates that all four samples of this class were misclassified, likely as the visually similar Corn_leaf_blight, revealing a critical point of confusion for corn diseases.
- **Severe confusion within potato diseases**: The second major area of weakness lies in distinguishing between potato diseases. The Potato_leaf_early_blight class performs very poorly, with a recall of only 14.3%, and this is compounded by the poor performance on Potato_leaf_late_blight (37.5% recall). The confusion between these two notoriously similar diseases is a primary source of error for the model, leading to a bidirectional misclassification, as they are very difficult to distinguish visually at certain stages.
- **Confusion within tomato diseases**: Several tomato diseases also present a significant challenge. Tomato_leaf_mosaic_virus is frequently misidentified, resulting in a low recall of 40.0%; it is most often mistaken for the visually similar Tomato_leaf_yellow_virus. Similarly, Tomato_leaf_bacterial_spot shows a modest recall score of 55.6%, as it is frequently confused with Tomato_leaf_late_blight. The Tomato_leaf (healthy) class also has a low recall of 50.0%, indicating frequent confusion with other tomato-related classes.

The detailed per-class performance for this optimized ensemble, including precision, recall, and F1-score for each of the 27 classes, is available in the full Classification Report in Appendix (Table A2).

Summary of the analysis: There is a clear pattern where the model excels at identifying unique, visually distinct classes but struggles significantly when faced with multiple diseases on the same plant species that share similar symptoms (e.g., spots, blights, or viral effects). The complete failure on Corn_Gray_leaf_spot and the severe confusion between potato blights are the most critical weaknesses.

4.4.2 Qualitative validation of failure cases

To concretely illustrate these areas of confusion and validate our central hypothesis—that the errors stem from intrinsic visual ambiguity—we performed a qualitative analysis of key failure cases. Figure 6 presents two representative examples of misclassified images, supplemented with their predicted probability distributions and Grad-CAM saliency maps. These maps highlight the image regions the model used to make its prediction.

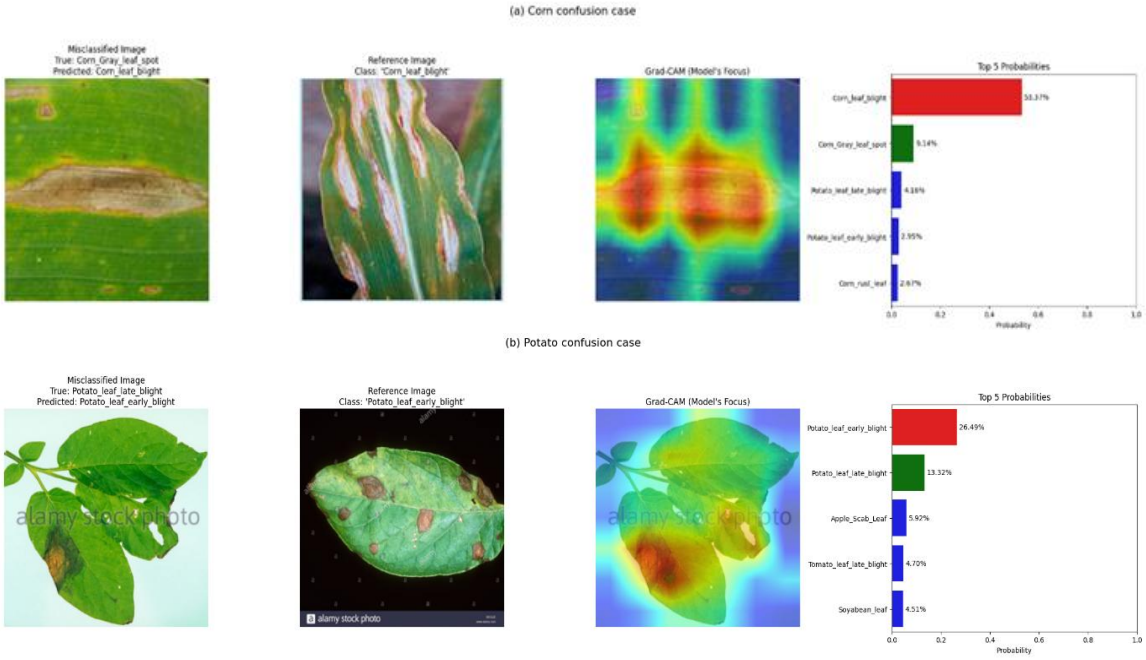


Figure 6: Qualitative analysis of classification errors. For each example, we show (1) the misclassified image, (2) a reference image of the predicted class, (3) the Grad-CAM saliency map showing the model's focus, and (4) the top predicted probabilities.

Examination of the visual examples in Figure 6 confirms and deepens the findings from the confusion matrix:

- **Ambiguity in corn diseases** (Figure 6a): This example shows a Corn_Gray_leaf_spot image misclassified as Corn_leaf_blight. The Grad-CAM map is particularly revealing: it shows that the model correctly focuses its attention on the large, necrotic lesion, confirming it has learned to identify relevant pathological features. The similarity of the necrotic lesions between the two corn diseases makes the model's confusion understandable, even for a human observer. Furthermore, the probability distribution shows only a moderate confidence of 53.37% for the wrong class, with the true class being the second most likely hypothesis. This demonstrates that while the model identifies the correct visual cues, the features of the lesion itself (shape, color, texture) are not sufficiently distinct for unambiguous classification.
- **Challenge of potato blights** (Figure 6b): The confusion between Potato_leaf_late_blight (True) and Potato_leaf_early_blight (Predicted) is notoriously difficult. More significantly, this example shows that distinguishing between early and late potato blight remains highly challenging, even on sharp images. The Grad-CAM map again indicates that the model is correctly analyzing the lesion area on the leaf. The very low prediction confidence (26.49%) and the high probability assigned to the true class (13.32%, the second highest) reveal the model's profound uncertainty.

These cases demonstrate that the model's errors are less due to learning defects than to the intrinsic complexity and ambiguity of the visual diagnosis task

4.4.3 Species-level performance evaluation

While fine-grained disease classification is challenging, another clinically relevant metric is the model's ability to correctly identify the plant species, irrespective of its health status. To assess this, we grouped the classes by species (e.g., all "Tomato" classes into a single "Tomato" category) and re-evaluated the optimized ensemble's performance. The results, presented in Table 8, show a dramatic improvement in performance.

The model achieves a macro-averaged F1-score of 0.9164 at the species level, a significant increase from the 0.7503 achieved for the fine-grained disease classification. This high level of performance is consistent across most species. Notably, the model achieves a perfect F1-score of 1.0000 for Corn, Squash, Strawberry, and Grape, indicating flawless species identification for these categories. Other major species like Apple (0.9180 F1-score) and Tomato (0.9308 F1-score) also demonstrate very high performance. The only notable exception is Cherry, which shows a lower recall (60%). Closer inspection reveals that these errors are primarily due to confusion with the related Peach species, likely owing to their visual similarity.

Overall, the high accuracy of 0.9246 at the species level confirms that the model robustly learns the general visual characteristics of each plant.

Table 8: Species-level classification report

	Precision	Recall	F1-score	Support
Apple	0.8750	0.9655	0.9180	29
Bell	0.8421	0.9412	0.8889	17
Blueberry	0.9091	0.9091	0.9091	11
Cherry	1.0000	0.6000	0.7500	10
Corn	1.0000	1.0000	1.0000	26
Peach	0.8889	0.8889	0.8889	9
Potato	0.8571	0.8182	0.8372	22
Raspberry	0.8750	1.0000	0.9333	7
Soyabean	1.0000	0.7500	0.8571	8
Squash	1.0000	1.0000	1.0000	6
Strawberry	1.0000	1.0000	1.0000	8
Tomato	0.9250	0.9367	0.9308	79
grape	1.0000	1.0000	1.0000	20
accuracy			0.9246	252
macro avg	0.9363	0.9084	0.9164	252
weighted avg	0.9276	0.9246	0.9229	252

Synthesis of the error analysis

Our multi-level error analysis leads to a unified conclusion: the model's errors are not random but are highly systematic and stem from the intrinsic difficulty of the task. The quantitative results show that confusions are concentrated on pairs of classes where visual distinction is intrinsically difficult, even for the human eye—for example, between different types of "blight," "spot," or "virus" on the same plant. The qualitative validation using Grad-CAM confirms this, demonstrating that the model correctly focuses on pathological regions but is hampered by ambiguous visual cues. Finally, the excellent performance at the species level reinforces the conclusion that the model has successfully learned high-level, species-specific features. The primary challenge, therefore, lies not in species identification, but in the fine-grained visual diagnosis of pathologies with similar manifestations. Improving discrimination in these complex cases constitutes a priority avenue for future work, potentially by integrating contextual information or using more targeted attention techniques.

5 Discussion

To contextualize our results within the existing research, we compare them to other studies that have also used the PlantDoc dataset. It is important to note that a direct, one-to-one comparison is challenging, as these studies may employ different data splits, preprocessing techniques, or evaluation protocols. This comparison, summarized in Table 9 (reporting Accuracy, as it is the only metric consistently available across these studies), should therefore be viewed as an indication of relative performance rather than a formal benchmark.

With this caveat in mind, our approach demonstrates highly competitive results. Our best individual model, the Swin-Base, achieves an accuracy of 0.7460. The primary finding, however, is that our optimized ensemble, obtained through a systematic ablation study, reaches a Macro F1-score of 0.7503 and an accuracy of 0.7619. This result compares favorably to previously published ensemble methods on this dataset, such as the stacking approach of R. Ramaprasad & S. Raman (0.7357) and the Snapshot Ensemble approach of Puangsuwan & Surinta (0.6951). Crucially, this strong performance is not achieved by developing a more complex aggregation mechanism; rather, it stems from a systematic and principled selection of a diverse set of base models. Our work demonstrates that the careful curation of an ensemble's composition—particularly by combining complementary CNNs and Transformers and pruning for predictive redundancy—is a more critical factor for success than the aggregation method itself.

Table 9: Comparison of results with other studies on the PlantDoc dataset.

Study	Method	Dataset	Accuracy
This study	Swin-Base	PlantDoc	0.7460
This study (Ensemble)	Ensemble (Swin-Base + ViT-Base + EfficientNetV2-B3)	PlantDoc	0.7619
R. Ramaprasad & S. Raman [33]	Stacking Ensemble (SEMFD-Net)	PlantDoc	0.7357
Menon et al. [34]	MobileNet-V2	PlantDoc	0.6980
Puangsuwan & Surinta [35]	Snapshot Ensemble (DenseNet201)	PlantDoc	0.6951
Chandra et al. [9]	SVM with siamese network	PlantDoc	0.6276
Moupoujou et al. [36]	MobileNet	PlantDoc	0.6014

Beyond a simple comparison of scores, our main contribution lies in the ensemble construction methodology. Existing literature has primarily explored two avenues for improving ensemble performance:

- Simple Voting Aggregation: Works such as those by Chaudhary et al. [15] or Mathew et al. [13] have shown that majority voting improved accuracy. Our study builds upon this by demonstrating the limitation of such approaches: the simple accumulation of models, even when aggregated with a more nuanced soft voting method, can be suboptimal if they are predictively redundant.

- Optimizing the Aggregation Strategy: To address the limitations of simple voting, a common approach is to use more advanced aggregation methods. However, our benchmark of such strategies in Section 4.2.2 provides a compelling counter-narrative. We found that various weighted voting schemes, even after probability calibration, offered only marginal gains or no improvement at all over

simple averaging. More importantly, even a sophisticated, two-level learning method like Stacking actually led to a performance degradation (Macro F1: 0.7201, Accuracy: 0.7222), significantly underperforming our pruned three-model ensemble (Macro F1: 0.7503, Accuracy: 0.7619). This empirically demonstrates a crucial principle: when predictive redundancy is the limiting factor, it is more effective to physically remove a redundant model (via ablation) than to attempt to mitigate its influence with complex aggregation logic or a meta-learner, which often struggle with data scarcity in such tasks. The strategic selection of an ensemble's composition is therefore the dominant optimization step.

Our work thus builds upon the foundational work of Mohanty et al. and Too et al., [3], [4] which established the superiority of Deep Learning on controlled, lab-condition datasets like PlantVillage. We extend their findings in two critical ways: first, by tackling the more challenging 'in-the-wild' conditions of the PlantDoc dataset, and second, by going beyond individual model performance to analyze their complex interactions within an ensemble—a crucial issue with the emergence of modern Transformer architectures.

Practical implications

Beyond the academic performance metrics, our findings offer significant practical implications for the development and deployment of real-world plant disease detection systems. The central conclusion—that a smaller, strategically curated ensemble can outperform a larger, more computationally expensive one—provides a direct guideline for practitioners.

For developers of agricultural technology, this means that instead of defaulting to the largest possible ensemble of models, a more resource-efficient and effective approach is to conduct a preliminary ablation study. By identifying and removing redundant or underperforming models, practitioners can achieve better overall performance while simultaneously reducing the computational overhead (memory, processing time) required for inference. This is particularly crucial for applications designed to run on edge devices with limited resources, such as smartphones, drones, or in-field sensors.

Consequently, our methodology translates into a tangible pathway for creating diagnostic tools that are not only more accurate but also faster and more accessible, thereby facilitating earlier and more widespread disease detection in agricultural practices.

Limitations and future work

Although our results establish a highly competitive performance and a robust methodology, it is important to acknowledge the limitations of this study to guide future work.

The primary limitation, is that our analysis, while diverse in terms of architectures, remains entirely based on visual features. Consequently, the errors of our best ensemble are concentrated on diseases with nearly identical symptoms, where visual diagnosis alone reaches its intrinsic limits. A second limitation is that, although PlantDoc contains real-world images, validating the robustness of our ensemble on broader data distributions—from different geographical regions or cultivation conditions—remains a necessary step for large-scale application.

These limitations open clear and promising avenues for future research, which can be categorized into three main directions.

The first direction involves exploring more advanced classification and ensemble strategies within the visual domain. Our current optimization was based on a systematic 'leave-one-out' ablation; an interesting alternative would be to compare this with constructive heuristics, such as a greedy forward selection algorithm. Additionally, building upon our finding that species identification is robust, a hierarchical classification approach (first identify the species, then the disease) could allow for smaller, specialized models to better tackle the subtle intra-species confusions.

The second direction, which addresses the intrinsic limits of purely visual diagnosis highlighted by our analysis, lies in the development of multimodal models. This is arguably the most promising path. Integrating structured metadata—such as crop variety, growth stage, or environmental sensor data (e.g., humidity, temperature)—could provide the critical discriminant context that is missing from the images alone. This would shift the paradigm from simple visual recognition to a more holistic, data-driven diagnostic system.

Finally, the third direction directly addresses the challenge of real-world generalization. Future work should focus on Transfer Learning and Domain Adaptation techniques. These methods would ensure that the performance of our optimized ensemble is maintained when faced with data from new field conditions without requiring extensive re-labeling, thereby guaranteeing the model's practical robustness.

6 Conclusion

In this study, we addressed the challenge of accurate plant disease detection by proposing a rigorous methodology for the optimization of Deep Learning ensembles. Through an advanced training protocol, systematically applied across multiple random seeds, we first established the robust individual performances of four state-of-the-art architectures.

Our central finding is that a pruned, three-model ensemble, identified through a systematic ablation study, achieves a highly competitive Macro F1-score of 0.7503 and an accuracy of 0.7619 on the complex PlantDoc dataset. This result surpasses the performance of the full four-model ensemble, a phenomenon we explain through correlation and diversity analysis, which revealed significant predictive redundancy. Our work empirically demonstrates that the naive aggregation of models is a suboptimal strategy. We have shown that carefully selecting an ensemble's composition to maximize predictive complementarity is a more effective

optimization lever than employing more complex aggregation methods like weighted voting or stacking, which proved less effective than our pruning approach.

While our results establish a robust performance benchmark, they also open clear avenues for future research, primarily through the integration of multimodal data to resolve the most ambiguous visual cases. In summary, our work positions the methodological pairing of ablation study and correlation analysis as an essential and pragmatic tool for designing robust and high-performing model ensembles in the field of precision agriculture.

References

[1] Padol, P. B. and Yadav, A. A. (2016). SVM classifier based grape leaf disease detection. Conference on Advances in Signal Processing (CASP), Pune, India, 2016, pp. 175-179. https://doi.org/10.1109/CASP.2016.7746160

[2] Vaishnnave, M. P., Devi, K. S., Srinivasan, P. and Jothi, G. A. P. (2019). Detection and Classification of Groundnut Leaf Diseases using KNN classifier, IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2019, pp. 1-5. https://doi.org/10.1109/ICSCAN.2019.8878733

[3] Mohanty, S. P., Hughes, D. P. and Salathe, M. (2016). Using deep learning for image-based plant disease detection. Front. Plant Sci. https://doi.org/10.3389/fpls.2016.01419

[4] Too, E. C., Yujian, L., Njuki, S., and Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. Comput. Electron. Agric. 161, pp. 272–279. https://doi.org/10.1016/j.compag.2018.03.032

[5] Hasan, R.I.; Yusuf, S.M.; Alzubaidi, L. (2020). Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion. Plants, 9, 1302. https://doi.org/10.3390/plants9101302

[6] Shoaib, M., Shah, B., Ei-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., Ali, F. (2023). An advanced deep learning models-based plant disease detection. A review of recent research. Front. Plant Sci., 14, 1158933. https://doi.org/10.3389/fpls.2023.1158933

[7] Akbar, M., Ullah, M., Shah, B., Khan, R.U., Hussain, T., Ali, F., Alenezi, F., Syed, I., Kwak, K.S. (2022). An effective deep learning approach for the classification of Bacteriosis in peach leave. Front. Plant Sci., 13, 1064854. https://doi.org/10.3389/fpls.2022.1064854

[8] Zhao, Y., Sun, C., Xu, X., and Chen, J. (2022). RIC-net: A plant disease classification model based on the fusion of inception and residual structure and embedded attention mechanism, C*omputers and Electronics in Agriculture* 193: 106644. https://doi.org/10.1016/j.compag.2021.106644

[9] Chandra, M., Redkar, S., Roy, S., Patil, P. (2020). Classification of Various Plant Diseases Using Deep Siamese Network. Available online: https://www.researchgate.net/publication/341322315 (accessed on 18 December 2025).

[10] Iftikhar, M., Kandhro, I. A., Kausar, N., Kehar A., Uddin, M. and Dandoush A. (2024). Plant disease management: A fine-tuned enhanced CNN approach with mobile app integration for early detection and classification, *Artif. Intell. Rev.*, vol. 57, no. 7, pp. 1-29. https://doi.org/10.1007/s10462-024-10809-z

[11] Ganaie, M., Hu, M., Malik, A., Tanveer, M., and Suganthan, P. (2022). Ensemble deep learning: A review. Eng. Appl. Artif. Intelligence, 115, 105151. https://doi.org/10.1016/j.engappai.2022.105151

[12] Li, H., Jin, Y., Zhong, J., and Zhao, R. (2021). A fruit tree disease diagnosis model based on stacking ensemble learning. Complexity, 2021, 6868592. https://doi.org/10.1155/2021/6868592

[13] Mathew, A., Antony, A., Mahadeshwar, Y., Khan, T., and Kulkarni, A. (2022). Plant disease detection using GLCM feature extractor and voting classification approach. Mat. Today, Proc. 58, Part 1, pp.407–415. https://doi.org/10.1016/j.matpr.2022.02.350

[14] Palanisamy, S., and Sanjana, N. (2023). Corn leaf disease detection using genetic algorithm and weighted voting, Proceedings of the 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India. (ICAECA). pp. 1-6. https://doi.org/10.1109/ICAECA56562.2023.10200196

[15] Chaudhary, A., Thakur, R., Kolhe, S., and Kamal, R. (2020). A particle swarm optimization-based ensemble for vegetable crop disease recognition. Comput. Electron. Agricult., 178, 105747. https://doi.org/10.1016/j.compag.2020.105747

[16] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249. https://doi.org/10.1002/widm.1249

[17] Giacinto, G., and Roli, F. (2001). Dynamic classifier selection based on multiple classifier behaviour. Pattern Recognition, 34(9), pp.1879-1881. https://doi.org/10.1016/S0031-3203(00)00150-3

[18] Tsoumakas, G., Katakis, I. (2007). Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM), 3(3), pp. 1-13. https://doi.org/10.4018/jdwm.2007070101

[19] Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A. (2004). Ensemble selection from libraries of models. Proceedings of the 21st International Conference on Machine Learning (ICML '04). https://doi.org/10.1145/1015330.1015432

[20] Astani, M., Hasheminejad, M. and Vaghefi, M. (2022). A diverse ensemble classifier for tomato disease recognition, *Comput. Electron. Agricult.*, vol. 198, Art. no. 107054. https://doi.org/10.1016/j.compag.2022.107054

[21] Shafik, W., Tufail, A. De Silva Liyanage, C. and R. Apong, A. A. H. M. (2024). Using transfer learning-

based plant disease classification and detection for sustainable agriculture, *BMC Plant Biol.*, vol. 24, no. 1, p. 136. https://doi.org/10.1186/s12870-024-04825-y

[22] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2010.11929

[23] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV48922.2021.00986

[24] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., Xie, S. (2022). A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). .https://doi.org/10.48550/arXiv.2201.03545

[25] Chen, J., Zhang, D. Zeb, A. and Nanehkaran, Y. A. (2021). Identification of Rice plant diseases using lightweight attention networks, *Expert Syst. Appl.*, vol. 169, Art. no. 114514. https://doi.org/10.1016/j.eswa.2020.114514

[26] Ouamane, A., Chouchane, A., Himeur, Y., Miniaoui, S., Atalla, S., Mansoor, W. (2025). Optimized Vision Transformers for Superior Plant Disease Detection, *IEEE Access*, vol. 13, pp. 48552-48570. https://doi.org/10.1109/ACCESS.2025.3547416

[27] Singh Thakur, P., Khanna, P., Sheorey, T., and Ojha, A. (2022). Explainable vision transformer enabled convolutional neural network for plant disease identification: Plantxvit, *arXiv:2207.07919*. https://doi.org/10.48550/arXiv.2207.07919

[28] Tabbakh, A. and Barpanda, S. S. (2023). A deep features extraction model based on the transfer learning model and vision transformer 'TLMViT' for plant, *IEEE Access*, vol. 11, pp. 45377-45392. https://doi.org/10.1109/ACCESS.2023.3273317

[29] Singh, D., Padgett, E., & T. S. G., A. K. (2020). PlantDoc: A Dataset for Visual Plant Disease Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.48550/arXiv.1911.10317

[30] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2980-2988. https://doi.org/10.48550/arXiv.1708.02002

[31] Kasneci, G. and Kasneci, E. (2024). Enriching Tabular Data with Contextual LLM Embeddings: A Comprehensive Ablation Study for Ensemble Classifiers, arXiv preprint arXiv:2411.01645. https://doi.org/10.48550/arXiv.2411.01645

[32] Grandini M., Bagli E., Visani G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756. https://doi.org/10.48550/arXiv.2008.05756

[33] Ramaprasad, R., Raman, S. (2022). SEMFD-Net: A Stacked Ensemble for Multiple Foliar Disease Classification. Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD) pp. 241-245. https://doi.org/10.1145/3493700.3493719

[34] Menon, V.; Ashwin, V.; Deepa, R.K. (2021). Plant disease detection using CNN and transfer learning. In Proceedings of the International Conference on Communication, Control and Information Sciences (ICCISc), Idukki, India, pp. 16–18; IEEE: Piscataway, NJ, USA, 2021; Volume 1, pp. 1–6. https://doi.org/10.1109/ICCISc52257.2021.9484957

[35] Puangsuwan T.; Surinta, O. (2021). Enhancement of plant leaf disease classification based on snapshot ensemble convolutional neural network. ICIC Exp Lett., 15(6), pp. 669–680. https://doi.org/10.24507/icicel.15.06.669

[36] Moupojou, E., Tagne, A., Retraint, F., Tadonkemwa, A., Dongmo, W., Tapamo, H., Nkenlifack, M. (2023). FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning. IEEE Access, 11, pp. 35398–35410. https://doi.org/10.1109/ACCESS.2023.3263042

Appendix A

Table A1: Class-wise Data Split (Stratification)

	Classe	**Total (Original Train)**	**Entraînement (85%)**	**Validation (15%)**	**Test (Officiel)**
0	Apple_Scab_Leaf	83	71	12	10
1	Apple_leaf	79	67	12	9
2	Apple_rust_leaf	96	82	14	10
3	Bell_pepper_leaf	34	29	5	8
4	Bell_pepper_leaf_spot	74	63	11	9
5	Blueberry_leaf	106	90	16	11
6	Cherry_leaf	47	40	7	10
7	Corn_Gray_leaf_spot	63	54	9	4
8	Corn_leaf_blight	182	155	27	12
9	Corn_rust_leaf	107	91	16	10
10	Peach_leaf	103	88	15	9
11	Potato_leaf_early_blight	157	133	24	14
12	Potato_leaf_late_blight	200	170	30	8
13	Raspberry_leaf	112	95	17	7
14	Soyabean_leaf	57	48	9	8
15	Squash_Powdery_mildew_leaf	124	105	19	6
16	Strawberry_leaf	88	75	13	8
17	Tomato_Early_blight_leaf	79	67	12	9
18	Tomato_Septoria_leaf_spot	145	123	22	12
19	Tomato_leaf	44	37	7	8
20	Tomato_leaf_bacterial_spot	101	86	15	9
21	Tomato_leaf_late_blight	101	86	15	10
22	Tomato_leaf_mosaic_virus	44	37	7	10
23	Tomato_leaf_yellow_virus	223	189	34	15
24	Tomato_mold_leaf	85	72	13	6
25	grape_leaf	63	54	9	12
26	grape_leaf_black_rot	71	60	11	8
	TOTAL	2668	2267	401	252

Table A2: Classification Report for the Ensemble without Convnext

Class	Precision	Recall	F1-Score	Support
Apple_Scab_Leaf	1.0000	0.9000	0.9474	10
Apple_leaf	0.6154	0.8889	0.7273	9
Apple_rust_leaf	0.9000	0.9000	0.9000	10
Bell_pepper_leaf	0.7778	0.8750	0.8235	8
Bell_pepper_leaf_spot	0.8000	0.8889	0.8421	9
Blueberry_leaf	0.9091	0.9091	0.9091	11
Cherry_leaf	1.0000	0.6000	0.7500	10
Corn_Gray_leaf_spot	0.0000	0.0000	0.0000	4
Corn_leaf_blight	0.6364	0.5833	0.6087	12
Corn_rust_leaf	1.0000	1.0000	1.0000	10
Peach_leaf	0.8889	0.8889	0.8889	9
Potato_leaf_early_blight	0.2500	0.1429	0.1818	14
Potato_leaf_late_blight	0.2308	0.3750	0.2857	8
Raspberry_leaf	0.8750	1.0000	0.9333	7
Soyabean_leaf	1.0000	0.7500	0.8571	8
Squash_Powdery_mildew_leaf	1.0000	1.0000	1.0000	6
Strawberry_leaf	1.0000	1.0000	1.0000	8
Tomato_Early_blight_leaf	0.6667	0.8889	0.7619	9
Tomato_Septoria_leaf_spot	0.7857	0.9167	0.8462	12
Tomato_leaf	0.8000	0.5000	0.6154	8
Tomato_leaf_bacterial_spot	0.6250	0.5556	0.5882	9
Tomato_leaf_late_blight	0.8000	0.8000	0.8000	10
Tomato_leaf_mosaic_virus	0.6667	0.4000	0.5000	10
Tomato_leaf_yellow_virus	0.7368	0.9333	0.8235	15
Tomato_mold_leaf	0.6667	0.6667	0.6667	6
grape_leaf	1.0000	1.0000	1.0000	12
grape_leaf_black_rot	1.0000	1.0000	1.0000	8
accuracy			0.7619	252
macro avg	0.7641	0.7542	0.7503	252
weighted avg	0.7708	0.7619	0.7571	252

Table A3: Classification Report for the Full Ensemble (Uncalibrated Soft Voting)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Apple_Scab_Leaf	0.8889	0.8000	0.8421	10
Apple_leaf	0.6667	0.8889	0.7619	9
Apple_rust_leaf	0.8182	0.9000	0.8571	10
Bell_pepper_leaf	0.7778	0.8750	0.8235	8
Bell_pepper_leaf_spot	0.7273	0.8889	0.8000	9
Blueberry_leaf	0.9091	0.9091	0.9091	11
Cherry_leaf	1.0000	0.7000	0.8235	10
Corn_Gray_leaf_spot	0.0000	0.0000	0.0000	4
Corn_leaf_blight	0.5455	0.5000	0.5217	12
Corn_rust_leaf	1.0000	0.9000	0.9474	10
Peach_leaf	1.0000	0.8889	0.9412	9
Potato_leaf_early_blight	0.2500	0.1429	0.1818	14
Potato_leaf_late_blight	0.2308	0.3750	0.2857	8
Raspberry_leaf	0.8750	1.0000	0.9333	7
Soyabean_leaf	1.0000	0.7500	0.8571	8
Squash_Powdery_mildew_leaf	1.0000	1.0000	1.0000	6
Strawberry_leaf	1.0000	1.0000	1.0000	8
Tomato_Early_blight_leaf	0.6154	0.8889	0.7273	9
Tomato_Septoria_leaf_spot	0.7333	0.9167	0.8148	12
Tomato_leaf	0.8000	0.5000	0.6154	8
Tomato_leaf_bacterial_spot	0.8000	0.4444	0.5714	9
Tomato_leaf_late_blight	0.8000	0.8000	0.8000	10
Tomato_leaf_mosaic_virus	0.6667	0.4000	0.5000	10
Tomato_leaf_yellow_virus	0.7368	0.9333	0.8235	15
Tomato_mold_leaf	0.6667	0.6667	0.6667	6
grape_leaf	1.0000	1.0000	1.0000	12
grape_leaf_black_rot	1.0000	1.0000	1.0000	8
accuracy			0.7500	252
macro avg	0.7596	0.7433	0.7409	252
weighted avg	0.7639	0.7500	0.7462	252

Table A4: Classification Report for the Full Ensemble (Calibrated Soft Voting)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Apple_Scab_Leaf	0.8889	0.8000	0.8421	10
Apple_leaf	0.6667	0.8889	0.7619	9
Apple_rust_leaf	0.8182	0.9000	0.8571	10
Bell_pepper_leaf	0.7778	0.8750	0.8235	8
Bell_pepper_leaf_spot	0.7273	0.8889	0.8000	9
Blueberry_leaf	0.9091	0.9091	0.9091	11
Cherry_leaf	1.0000	0.6000	0.7500	10
Corn_Gray_leaf_spot	0.0000	0.0000	0.0000	4
Corn_leaf_blight	0.5455	0.5000	0.5217	12
Corn_rust_leaf	1.0000	0.9000	0.9474	10
Peach_leaf	0.8889	0.8889	0.8889	9
Potato_leaf_early_blight	0.2500	0.1429	0.1818	14
Potato_leaf_late_blight	0.2308	0.3750	0.2857	8
Raspberry_leaf	0.8750	1.0000	0.9333	7
Soyabean_leaf	1.0000	0.7500	0.8571	8
Squash_Powdery_mildew_leaf	1.0000	1.0000	1.0000	6
Strawberry_leaf	1.0000	1.0000	1.0000	8
Tomato_Early_blight_leaf	0.6154	0.8889	0.7273	9
Tomato_Septoria_leaf_spot	0.7143	0.8333	0.7692	12
Tomato_leaf	0.8000	0.5000	0.6154	8
Tomato_leaf_bacterial_spot	0.6667	0.4444	0.5333	9
Tomato_leaf_late_blight	0.8000	0.8000	0.8000	10
Tomato_leaf_mosaic_virus	0.7143	0.5000	0.5882	10
Tomato_leaf_yellow_virus	0.7368	0.9333	0.8235	15
Tomato_mold_leaf	0.8000	0.6667	0.7273	6
grape_leaf	1.0000	1.0000	1.0000	12
grape_leaf_black_rot	1.0000	1.0000	1.0000	8
accuracy			0.7460	252
macro avg	0.7565	0.7402	0.7387	252
weighted avg	0.7594	0.7460	0.7428	252

Table A5: Classification Report for the full Ensemble (Calibrated Weighted (F1))

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Apple_Scab_Leaf	0.8889	0.8000	0.8421	10
Apple_leaf	0.6667	0.8889	0.7619	9
Apple_rust_leaf	0.8182	0.9000	0.8571	10
Bell_pepper_leaf	0.7778	0.8750	0.8235	8
Bell_pepper_leaf_spot	0.7273	0.8889	0.8000	9
Blueberry_leaf	0.9091	0.9091	0.9091	11
Cherry_leaf	1.0000	0.6000	0.7500	10
Corn_Gray_leaf_spot	0.0000	0.0000	0.0000	4
Corn_leaf_blight	0.5455	0.5000	0.5217	12
Corn_rust_leaf	1.0000	0.9000	0.9474	10
Peach_leaf	0.8889	0.8889	0.8889	9
Potato_leaf_early_blight	0.2500	0.1429	0.1818	14
Potato_leaf_late_blight	0.2308	0.3750	0.2857	8
Raspberry_leaf	0.8750	1.0000	0.9333	7
Soyabean_leaf	1.0000	0.7500	0.8571	8
Squash_Powdery_mildew_leaf	1.0000	1.0000	1.0000	6
Strawberry_leaf	1.0000	1.0000	1.0000	8
Tomato_Early_blight_leaf	0.6154	0.8889	0.7273	9
Tomato_Septoria_leaf_spot	0.7333	0.9167	0.8148	12
Tomato_leaf	0.8000	0.5000	0.6154	8
Tomato_leaf_bacterial_spot	0.8000	0.4444	0.5714	9
Tomato_leaf_late_blight	0.8000	0.8000	0.8000	10
Tomato_leaf_mosaic_virus	0.7143	0.5000	0.5882	10
Tomato_leaf_yellow_virus	0.7368	0.9333	0.8235	15
Tomato_mold_leaf	0.8000	0.6667	0.7273	6
grape_leaf	1.0000	1.0000	1.0000	12
grape_leaf_black_rot	1.0000	1.0000	1.0000	8
accuracy			0.7500	252
macro avg	0.7621	0.7433	0.7418	252
weighted avg	0.7650	0.7500	0.7464	252

Table A6: Classification Report for the full Ensemble (Calibrated Weighted (LogLoss))

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Apple_Scab_Leaf	0.8889	0.8000	0.8421	10
Apple_leaf	0.6667	0.8889	0.7619	9
Apple_rust_leaf	0.8182	0.9000	0.8571	10
Bell_pepper_leaf	0.7778	0.8750	0.8235	8
Bell_pepper_leaf_spot	0.7273	0.8889	0.8000	9
Blueberry_leaf	0.9091	0.9091	0.9091	11
Cherry_leaf	1.0000	0.7000	0.8235	10
Corn_Gray_leaf_spot	0.0000	0.0000	0.0000	4
Corn_leaf_blight	0.5455	0.5000	0.5217	12
Corn_rust_leaf	1.0000	0.9000	0.9474	10
Peach_leaf	1.0000	0.8889	0.9412	9
Potato_leaf_early_blight	0.2500	0.1429	0.1818	14
Potato_leaf_late_blight	0.2308	0.3750	0.2857	8
Raspberry_leaf	0.8750	1.0000	0.9333	7
Soyabean_leaf	1.0000	0.7500	0.8571	8
Squash_Powdery_mildew_leaf	1.0000	1.0000	1.0000	6
Strawberry_leaf	1.0000	1.0000	1.0000	8
Tomato_Early_blight_leaf	0.6154	0.8889	0.7273	9
Tomato_Septoria_leaf_spot	0.7143	0.8333	0.7692	12
Tomato_leaf	1.0000	0.5000	0.6667	8
Tomato_leaf_bacterial_spot	0.6667	0.4444	0.5333	9
Tomato_leaf_late_blight	0.8000	0.8000	0.8000	10
Tomato_leaf_mosaic_virus	0.7500	0.6000	0.6667	10
Tomato_leaf_yellow_virus	0.7368	0.9333	0.8235	15
Tomato_mold_leaf	0.8000	0.6667	0.7273	6
grape_leaf	1.0000	1.0000	1.0000	12
grape_leaf_black_rot	1.0000	1.0000	1.0000	8
accuracy			0.7540	252
macro avg	0.7693	0.7476	0.7481	252
weighted avg	0.7711	0.7540	0.7524	252

Table A7: Classification Report for the Full Ensemble Stacking (Optimized XGBoost)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Apple_Scab_Leaf	0.9000	0.9000	0.9000	10
Apple_leaf	0.6364	0.7778	0.7000	9
Apple_rust_leaf	0.9091	1.0000	0.9524	10
Bell_pepper_leaf	0.7000	0.8750	0.7778	8
Bell_pepper_leaf_spot	0.8571	0.6667	0.7500	9
Blueberry_leaf	0.9000	0.8182	0.8571	11
Cherry_leaf	1.0000	0.7000	0.8235	10
Corn_Gray_leaf_spot	0.2500	0.2500	0.2500	4
Corn_leaf_blight	0.6667	0.6667	0.6667	12
Corn_rust_leaf	0.9000	0.9000	0.9000	10
Peach_leaf	1.0000	0.8889	0.9412	9
Potato_leaf_early_blight	0.5000	0.2857	0.3636	14
Potato_leaf_late_blight	0.2727	0.3750	0.3158	8
Raspberry_leaf	0.8750	1.0000	0.9333	7
Soyabean_leaf	1.0000	0.7500	0.8571	8
Squash_Powdery_mildew_leaf	1.0000	1.0000	1.0000	6
Strawberry_leaf	1.0000	1.0000	1.0000	8
Tomato_Early_blight_leaf	0.3636	0.4444	0.4000	9
Tomato_Septoria_leaf_spot	0.3889	0.5833	0.4667	12
Tomato_leaf	1.0000	0.3750	0.5455	8
Tomato_leaf_bacterial_spot	0.2727	0.3333	0.3000	9
Tomato_leaf_late_blight	0.5833	0.7000	0.6364	10
Tomato_leaf_mosaic_virus	0.7778	0.7000	0.7368	10
Tomato_leaf_yellow_virus	0.8235	0.9333	0.8750	15
Tomato_mold_leaf	0.7500	0.5000	0.6000	6
grape_leaf	0.9231	1.0000	0.9600	12
grape_leaf_black_rot	1.0000	0.8750	0.9333	8
accuracy			0.7222	252
macro avg	0.7500	0.7148	0.7201	252
weighted avg	0.7511	0.7222	0.7248	252

Appendix B: Test Set Filenames

To ensure full auditability and reproducibility of our results, this appendix lists the complete set of the 252 image filenames that constitute the test set used in all our experiments. The file paths are relative to the main dataset directory.

Table B1: Complete list of test set filenames with index

#	File Path	#	File Path
1	Apple_Scab_Leaf/test_Apple Scab Leaf_1.jpg	127	Raspberry_leaf/test_Raspberry leaf_1.jpg
2	Apple_Scab_Leaf/test_Apple Scab Leaf_10.jpg	128	Raspberry_leaf/test_Raspberry leaf_2.jpg
3	Apple_Scab_Leaf/test_Apple Scab Leaf_2.jpg	129	Raspberry_leaf/test_Raspberry leaf_3.jpg
4	Apple_Scab_Leaf/test_Apple Scab Leaf_3.jpg	130	Raspberry_leaf/test_Raspberry leaf_4.jpg
5	Apple_Scab_Leaf/test_Apple Scab Leaf_4.jpg	131	Raspberry_leaf/test_Raspberry leaf_5.jpg
6	Apple_Scab_Leaf/test_Apple Scab Leaf_5.jpg	132	Raspberry_leaf/test_Raspberry leaf_6.jpg
7	Apple_Scab_Leaf/test_Apple Scab Leaf_6.jpg	133	Raspberry_leaf/test_Raspberry leaf_7.jpg
8	Apple_Scab_Leaf/test_Apple Scab Leaf_7.jpg	134	Soyabean_leaf/test_Soyabean leaf_1.jpg
9	Apple_Scab_Leaf/test_Apple Scab Leaf_8.jpg	135	Soyabean_leaf/test_Soyabean leaf_2.jpg
10	Apple_Scab_Leaf/test_Apple Scab Leaf_9.jpg	136	Soyabean_leaf/test_Soyabean leaf_3.jpg
11	Apple_leaf/test_Apple leaf_1.jpg	137	Soyabean_leaf/test_Soyabean leaf_4.jpg
12	Apple_leaf/test_Apple leaf_2.jpg	138	Soyabean_leaf/test_Soyabean leaf_5.jpg
13	Apple_leaf/test_Apple leaf_3.jpg	139	Soyabean_leaf/test_Soyabean leaf_6.jpg
14	Apple_leaf/test_Apple leaf_4.jpg	140	Soyabean_leaf/test_Soyabean leaf_7.jpg
15	Apple_leaf/test_Apple leaf_5.jpg	141	Soyabean_leaf/test_Soyabean leaf_8.jpg
16	Apple_leaf/test_Apple leaf_6.jpg	142	Squash_Powdery_mildew_leaf/test_Squash Powdery mildew leaf_1.jpg
17	Apple_leaf/test_Apple leaf_7.jpg	143	Squash_Powdery_mildew_leaf/test_Squash Powdery mildew leaf_2.jpg
18	Apple_leaf/test_Apple leaf_8.jpg	144	Squash_Powdery_mildew_leaf/test_Squash Powdery mildew leaf_3.jpg
19	Apple_leaf/test_Apple leaf_9.jpg	145	Squash_Powdery_mildew_leaf/test_Squash Powdery mildew leaf_4.jpg
20	Apple_rust_leaf/test_Apple rust leaf_1.jpg	146	Squash_Powdery_mildew_leaf/test_Squash Powdery mildew leaf_5.jpg
21	Apple_rust_leaf/test_Apple rust leaf_10.jpg	147	Squash_Powdery_mildew_leaf/test_Squash Powdery mildew leaf_6.jpg
22	Apple_rust_leaf/test_Apple rust leaf_2.jpg	148	Strawberry_leaf/test_Strawberry leaf_1.jpg
23	Apple_rust_leaf/test_Apple rust leaf_3.jpg	149	Strawberry_leaf/test_Strawberry leaf_2.jpg
24	Apple_rust_leaf/test_Apple rust leaf_4.jpg	150	Strawberry_leaf/test_Strawberry leaf_3.jpg
25	Apple_rust_leaf/test_Apple rust leaf_5.jpg	151	Strawberry_leaf/test_Strawberry leaf_4.jpg
26	Apple_rust_leaf/test_Apple rust leaf_6.jpg	152	Strawberry_leaf/test_Strawberry leaf_5.jpg
27	Apple_rust_leaf/test_Apple rust leaf_7.jpg	153	Strawberry_leaf/test_Strawberry leaf_6.jpg
28	Apple_rust_leaf/test_Apple rust leaf_8.jpg	154	Strawberry_leaf/test_Strawberry leaf_7.jpg
29	Apple_rust_leaf/test_Apple rust leaf_9.jpg	155	Strawberry_leaf/test_Strawberry leaf_8.jpg
30	Bell_pepper_leaf/test_Bell_pepper leaf_1.jpg	156	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_1.jpg
31	Bell_pepper_leaf/test_Bell_pepper leaf_2.jpg	157	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_2.jpg
32	Bell_pepper_leaf/test_Bell_pepper leaf_3.jpg	158	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_3.jpg

#	**File Path**	#	**File Path**
33	Bell_pepper_leaf/test_Bell_pepper leaf_4.jpg	159	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_4.jpg
34	Bell_pepper_leaf/test_Bell_pepper leaf_5.jpg	160	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_5.jpg
35	Bell_pepper_leaf/test_Bell_pepper leaf_6.jpg	161	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_6.jpg
36	Bell_pepper_leaf/test_Bell_pepper leaf_7.jpg	162	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_7.jpg
37	Bell_pepper_leaf/test_Bell_pepper leaf_8.jpg	163	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_8.jpg
38	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_1.jpg	164	Tomato_Early_blight_leaf/test_Tomato Early blight leaf_9.jpg
39	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_2.jpg	165	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_10.jpg
40	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_3.jpg	166	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_11.jpg
41	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_4.jpg	167	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_1_1.jpg
42	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_5.jpg	168	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_1_2.jpg
43	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_6.jpg	169	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_2.jpg
44	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_7.jpg	170	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_3.jpg
45	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_8.jpg	171	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_4.jpg
46	Bell_pepper_leaf_spot/test_Bell_pepper leaf spot_9.jpg	172	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_5.jpg
47	Blueberry_leaf/test_Blueberry leaf_1.jpg	173	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_6.jpg
48	Blueberry_leaf/test_Blueberry leaf_10.jpg	174	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_7.jpg
49	Blueberry_leaf/test_Blueberry leaf_11.jpg	175	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_8.jpg
50	Blueberry_leaf/test_Blueberry leaf_2.jpg	176	Tomato_Septoria_leaf_spot/test_Tomato Septoria leaf spot_9.jpg
51	Blueberry_leaf/test_Blueberry leaf_3.jpg	177	Tomato_leaf/test_Tomato leaf_1.jpg
52	Blueberry_leaf/test_Blueberry leaf_4.jpg	178	Tomato_leaf/test_Tomato leaf_2.jpg
53	Blueberry_leaf/test_Blueberry leaf_5.jpg	179	Tomato_leaf/test_Tomato leaf_3.jpg
54	Blueberry_leaf/test_Blueberry leaf_6.jpg	180	Tomato_leaf/test_Tomato leaf_4.jpg
55	Blueberry_leaf/test_Blueberry leaf_7.jpg	181	Tomato_leaf/test_Tomato leaf_5.jpg
56	Blueberry_leaf/test_Blueberry leaf_8.jpg	182	Tomato_leaf/test_Tomato leaf_6.jpg
57	Blueberry_leaf/test_Blueberry leaf_9.jpg	183	Tomato_leaf/test_Tomato leaf_7.jpg
58	Cherry_leaf/test_Cherry leaf_1.jpg	184	Tomato_leaf/test_Tomato leaf_8.jpg
59	Cherry_leaf/test_Cherry leaf_10.jpg	185	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_1.jpg
60	Cherry_leaf/test_Cherry leaf_2.jpg	186	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_2.jpg
61	Cherry_leaf/test_Cherry leaf_3.jpg	187	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_3.jpg
62	Cherry_leaf/test_Cherry leaf_4.jpg	188	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_4.jpg
63	Cherry_leaf/test_Cherry leaf_5.jpg	189	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_5.jpg
64	Cherry_leaf/test_Cherry leaf_6.jpg	190	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_6.jpg

65	Cherry_leaf/test_Cherry leaf_7.jpg	191	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_7.jpg
#	**File Path**	#	**File Path**
66	Cherry_leaf/test_Cherry leaf_8.jpg	192	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_8.jpg
67	Cherry_leaf/test_Cherry leaf_9.jpg	193	Tomato_leaf_bacterial_spot/test_Tomato leaf bacterial spot_9.jpg
68	Corn_Gray_leaf_spot/test_Corn Gray leaf spot_1.jpg	194	Tomato_leaf_late_blight/test_Tomato leaf late blight_1.jpg
69	Corn_Gray_leaf_spot/test_Corn Gray leaf spot_2.jpg	195	Tomato_leaf_late_blight/test_Tomato leaf late blight_10.jpg
70	Corn_Gray_leaf_spot/test_Corn Gray leaf spot_3.jpg	196	Tomato_leaf_late_blight/test_Tomato leaf late blight_2.jpg
71	Corn_Gray_leaf_spot/test_Corn Gray leaf spot_4.jpg	197	Tomato_leaf_late_blight/test_Tomato leaf late blight_3.jpg
72	Corn_leaf_blight/test_Corn leaf blight_1.jpg	198	Tomato_leaf_late_blight/test_Tomato leaf late blight_4.jpg
73	Corn_leaf_blight/test_Corn leaf blight_10.jpg	199	Tomato_leaf_late_blight/test_Tomato leaf late blight_5.jpg
74	Corn_leaf_blight/test_Corn leaf blight_11.jpg	200	Tomato_leaf_late_blight/test_Tomato leaf late blight_6.jpg
75	Corn_leaf_blight/test_Corn leaf blight_12.jpg	201	Tomato_leaf_late_blight/test_Tomato leaf late blight_7.jpg
76	Corn_leaf_blight/test_Corn leaf blight_2.jpg	202	Tomato_leaf_late_blight/test_Tomato leaf late blight_8.jpg
77	Corn_leaf_blight/test_Corn leaf blight_3.jpg	203	Tomato_leaf_late_blight/test_Tomato leaf late blight_9.jpg
78	Corn_leaf_blight/test_Corn leaf blight_4.jpg	204	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_1.jpg
79	Corn_leaf_blight/test_Corn leaf blight_5.jpg	205	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_10.jpg
80	Corn_leaf_blight/test_Corn leaf blight_6.jpg	206	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_2.jpg
81	Corn_leaf_blight/test_Corn leaf blight_7.jpg	207	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_3.jpg
82	Corn_leaf_blight/test_Corn leaf blight_8.jpg	208	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_4.jpg
83	Corn_leaf_blight/test_Corn leaf blight_9.jpg	209	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_5.jpg
84	Corn_rust_leaf/test_Corn rust leaf_1.jpg	210	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_6.jpg
85	Corn_rust_leaf/test_Corn rust leaf_10.jpg	211	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_7.jpg
86	Corn_rust_leaf/test_Corn rust leaf_2.jpg	212	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_8.jpg
87	Corn_rust_leaf/test_Corn rust leaf_3.jpg	213	Tomato_leaf_mosaic_virus/test_Tomato leaf mosaic virus_9.jpg
88	Corn_rust_leaf/test_Corn rust leaf_4.jpg	214	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_1.jpg
89	Corn_rust_leaf/test_Corn rust leaf_5.jpg	215	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_2_1.jpg
90	Corn_rust_leaf/test_Corn rust leaf_6.jpg	216	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_2_2.jpg
91	Corn_rust_leaf/test_Corn rust leaf_7.jpg	217	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_2_3.jpg
92	Corn_rust_leaf/test_Corn rust leaf_8.jpg	218	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_2_4.jpg
93	Corn_rust_leaf/test_Corn rust leaf_9.jpg	219	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_2_5.jpg
94	Peach_leaf/test_Peach leaf_1.jpg	220	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_2_6.jpg
95	Peach_leaf/test_Peach leaf_2.jpg	121	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_3_1.jpg
#	**File Path**	#	**File Path**

96	Peach_leaf/test_Peach leaf_3.jpg	222	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_3_1.jpg
97	Peach_leaf/test_Peach leaf_4.jpg	223	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_4_1.jpg
98	Peach_leaf/test_Peach leaf_5.jpg	224	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_4_2.jpg
99	Peach_leaf/test_Peach leaf_6.jpg	225	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_5.jpg
100	Peach_leaf/test_Peach leaf_7.jpg	226	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_6_1.jpg
101	Peach_leaf/test_Peach leaf_8.jpg	227	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_6_2.jpg
102	Peach_leaf/test_Peach leaf_9.jpg	228	Tomato_leaf_yellow_virus/test_Tomato leaf yellow virus_6_3.jpg
103	Potato_leaf_early_blight/test_Potato leaf early blight_1.jpg	229	Tomato_mold_leaf/test_Tomato mold leaf_1.jpg
104	Potato_leaf_early_blight/test_Potato leaf early blight_2.jpg	230	Tomato_mold_leaf/test_Tomato mold leaf_2.jpg
105	Potato_leaf_early_blight/test_Potato leaf early blight_3.jpg	231	Tomato_mold_leaf/test_Tomato mold leaf_3.jpg
106	Potato_leaf_early_blight/test_Potato leaf early blight_4_1.jpg	232	Tomato_mold_leaf/test_Tomato mold leaf_4.jpg
107	Potato_leaf_early_blight/test_Potato leaf early blight_4_2.jpg	233	Tomato_mold_leaf/test_Tomato mold leaf_5.jpg
108	Potato_leaf_early_blight/test_Potato leaf early blight_4_3.jpg	234	Tomato_mold_leaf/test_Tomato mold leaf_6.jpg
109	Potato_leaf_early_blight/test_Potato leaf early blight_4_4.jpg	235	grape_leaf/test_grape leaf_1.jpg
110	Potato_leaf_early_blight/test_Potato leaf early blight_5.jpg	236	grape_leaf/test_grape leaf_10.jpg
111	Potato_leaf_early_blight/test_Potato leaf early blight_6_1.jpg	237	grape_leaf/test_grape leaf_11.jpg
112	Potato_leaf_early_blight/test_Potato leaf early blight_6_2.jpg	238	grape_leaf/test_grape leaf_12.jpg
113	Potato_leaf_early_blight/test_Potato leaf early blight_6_3.jpg	239	grape_leaf/test_grape leaf_2.jpg
114	Potato_leaf_early_blight/test_Potato leaf early blight_6_4.jpg	240	grape_leaf/test_grape leaf_3.jpg
115	Potato_leaf_early_blight/test_Potato leaf early blight_7.jpg	241	grape_leaf/test_grape leaf_4.jpg
116	Potato_leaf_early_blight/test_Potato leaf early blight_8.jpg	242	grape_leaf/test_grape leaf_5.jpg
117	Potato_leaf_late_blight/test_Potato leaf late blight_1.jpg	243	grape_leaf/test_grape leaf_6.jpg
118	Potato_leaf_late_blight/test_Potato leaf late blight_2.jpg	244	grape_leaf/test_grape leaf_7.jpg
119	Potato_leaf_late_blight/test_Potato leaf late blight_3.jpg	245	grape_leaf/test_grape leaf_8.jpg
120	Potato_leaf_late_blight/test_Potato leaf late blight_4.jpg	246	grape_leaf/test_grape leaf_9.jpg
121	Potato_leaf_late_blight/test_Potato leaf late blight_5.jpg	247	grape_leaf_black_rot/test_grape leaf black rot_1.jpg
122	Potato_leaf_late_blight/test_Potato leaf late blight_6.jpg	248	grape_leaf_black_rot/test_grape leaf black rot_2.jpg
123	Potato_leaf_late_blight/test_Potato leaf late blight_7.jpg	249	grape_leaf_black_rot/test_grape leaf black rot_3.jpg
124	Potato_leaf_late_blight/test_Potato leaf late blight_8.jpg	250	grape_leaf_black_rot/test_grape leaf black rot_4.jpg
125	Raspberry_leaf/test_Raspberry leaf_1.jpg	251	grape_leaf_black_rot/test_grape leaf black rot_5.jpg
126	Raspberry_leaf/test_Raspberry leaf_2.jpg	252	grape_leaf_black_rot/test_grape leaf black rot_6.jpg

Real-Time Detection of Public Relations Crisis Signals Using Hybrid LSTM and Isolation Forest Model

Yuchao Zeng
School of Management, Zhanjiang University of Science and Technology, Zhanjiang 524094, China
E-mail: YuchaoZeng@outlook.com

Keywords: public relations crisis, long and short-term memory networks, isolation forest, real-time detection

Received: October 20, 2025

With the widespread use of social media, public relations crises have become increasingly rapid and unpredictable. Traditional manual monitoring and public opinion analysis methods are difficult to cope with the rapidly developing information environment. To address this challenge, this study proposes a new real-time PR crisis signal detection model that effectively combines long short-term memory networks and isolation forest algorithms. This model aims to achieve early warning of potential crises by efficiently processing large amounts of social media and news data. The experiment was conducted on a real-world public relations crisis event dataset, which includes text, user interaction, and sentiment analysis data. The performance of the proposed fusion model was compared with independent LSTM and isolated forest models as baselines. The experimental results show that the fusion model performs robustly in various types of public relations crisis signals. It achieved significant high accuracy in detecting social media popularity and public opinion fluctuations, improving by 3.4% and 6.3% respectively compared to single LSTM and Isolation Forest models. In addition, the fusion model exhibits significantly enhanced robustness in high noise environments. At a pollution rate of 60%, it maintained an accuracy of 54.1% and an F1 score of 0.823, demonstrating its powerful anomaly detection capability. The proposed fusion model provides an accurate, real-time, and robust solution for PR crisis signal detection by utilizing the complementary advantages of LSTM to capture time patterns and isolation forest to identify anomalies, which has significant practical application value for enterprises and organizations.

Povzetek: Prispevek predstavi realnočasovni model za zgodnje zaznavanje signalov PR-kriz, ki združuje LSTM in Isolation Forest ter pri analizi družbenih medijev doseže boljšo točnost in robustnost kot posamezni pristopi.

1 Introduction

With the rapid development of information technology, social media has become a key platform for people to communicate and express their opinions daily, and enterprises and organizations are faced with the risk of public relations crisis, which will affect the reputation of enterprises and may also cause serious economic losses [1]. It is crucial for enterprises and organizations to detect crisis signals in time and effectively and prevent them from occurring. Traditional crisis detection methods mainly rely on manual monitoring and event review, which are inefficient and easy to miss potential crisis signals.

Technology has developed quite rapidly. Automated monitoring systems relying on artificial intelligence have shown great potential in real-time detection of public relations crises. Long short-term memory networks are a type of neural network frequently used in time series data analysis, and they have achieved some results in tasks such as text classification and sentiment analysis [2]. Can learn the time series features in historical data, which can effectively identify the early signs of crisis signals and make predictions more accurate. However, using only one handle complex and changeable public relations crisis signals may have some shortcomings in such a noisy data environment.

To address this challenge, a fusion model that combines LSTM and the Isolation Forest algorithm has been developed. Isolation Forest relies on tree structures. This approach can effectively identify anomaly patterns from large-scale data, integrating the advantages of LSTM and Segregated Forest[3]. There is hope nature of public relations crisis signal detection. When confronted with a vast number of social media data, this fusion model can more accurately identify potential crisis events.

Research shows that the combination of LSTM and Isolation Forest results in many fields. In the aspect of financial risk detection, the approach of combining LSTM and segregated forests has been proven to be effective in predicting fluctuations and abnormal situations in the financial market. Here, relevant literature can be referred to [4]. The same situation also occurs in the sentiment analysis and crisis monitoring of social media data. Compared with traditional methods, the fusion model appears more stable and accurate. This article aims to explore the application of this method in the detection of public relations crisis signals, which can provide enterprises and organizations with a more

accurate and real-time monitoring crisis early warning system [5].

In order to systematically evaluate the performance of the LSTM and isolated forest fusion model, this study proposes three core issues: whether the fusion model can surpass the single model in crisis detection accuracy; How robust is it in high noise environments on social media; And whether it can meet the efficiency requirements of real-time computing. Based on this assumption, it is predicted that the fusion model will significantly improve detection accuracy and F1 score, maintain stable performance under over 50% noise pollution, and have throughput and latency characteristics suitable for real-time scenarios.

2 Theoretical basis and related research

2.1 Theory of LSTM and isolation forest fusion algorithm

Put it simply, it is a recurrent neural network. By introducing mechanisms such as input gates, forgetting gates, and output gates [6, 7], it has effectively addressed the gradient vanishing and explosion problems that occur in traditional RNNS when processing long sequence data. As a result, when LSTM processes time series data, this way, long-term dependencies can be captured. For real-time detection tasks of public relations crises, LSTM will analyze the time series patterns in historical data, identify possible crisis signals from them, and promptly issue early warnings to decision-makers. LSTM has a strong time series learning ability, which makes it of great application value when dealing with dynamic and changing crisis scenarios [8],

Isolation Forest algorithm is an anomaly detection method based on tree structure, and its core concept is to distinguish anomaly points from normal data points by randomly segmenting data sets [9, 10]. Isolation Forest relies on building multiple random trees to "isolate" data points. Outlier points generally require fewer segmentation steps and can be isolated earlier. Compared with traditional distance-based algorithms, Isolation Forest does not rely on data points. Direct distance calculation between points is more efficient, especially suitable for processing high-dimensional data and large-scale data sets [11]. In practical application, Isolation Forest can accurately detect sudden abnormal fluctuations in public relations crisis, provide strong support for real-time monitoring system,

In this paper, the combination of LSTM and Isolation Forest can give full play to their respective strengths and improve the accuracy and robustness of public relations crisis detection [12]. LSTM performs very well in learning and predicting the long-term dependencies of time series data. It can accurately capture the development trend of public relations events. Isolation Forest, on the other hand, is particularly adept at identifying outliers in the data. If the time series data shows abnormal fluctuations that are significantly different from historical patterns, Segregated Forest can immediately recognize them and issue early warnings for potential crises [13]. By combining LSTM with isolated forest, this algorithm can accurately predict time series patterns and promptly detect sudden abnormal changes. This can enhance the ability to recognize crisis signals and also make the response speed faster.

The integration of LSTM and isolated forest does have some significant advantages. However, when it comes to practical application, there are still some challenges. For LSTM to operate normally, it computing resources and data to support it. During its training process, it has relatively high requirements for real-time performance. As a result, It may have a certain degree of impact on the response speed of the system [14, 15]. Although isolated forests can handle high-dimensional data relatively effectively, when encountering data noise, they may result in false positives or false negatives. In future research, it is necessary to optimize the combination of these two algorithms to enhance the computational efficiency of the model and anomaly detection. Only in this way can we deal with the complex and ever-changing public relations crisis scenarios.

2.2 Current status of public relations crisis signal detection

Public relations crisis signal detection is a key topic in the field of public management and enterprise crisis management continuous popularity of social media, public relations crisis occurs rapidly and unpredictably. Traditional crisis detection methods mostly rely on expert experience and public opinion analysis, lacking automation and real-time. How to use big data analysis and machine learning algorithms to carry out real-time monitoring and early warning of public relations crisis signals has become a key issue in current research and practice [16, 17].

Public relations crisis signal detection methods are mainly based on content analysis, sentiment analysis, social network analysis and other technologies. These methods identify potential crisis signals by analyzing news reports, social media content, public opinion comments and other data. Sentiment analysis technology can extract emotional polarity from a large number of texts, judge the public's attitude and emotional fluctuation towards an event, and provide basis for crisis early warning. However, most of these traditional methods are limited to static data analysis, lacking in-depth excavation of dynamic changes and timeline relationships in the process of event development, resulting in slow response speed when emergencies occur [18].

Machine learning technology, more and more scholars and enterprises have begun to try to use more advanced algorithms public relations crisis detection. As a deep learning algorithm that can process time series data, LSTM shows strong ability in crisis signal detection. LSTM can learn potential time series patterns from a large amount of historical data and identify early signals of crisis events [19, 20]. However, when a single LSTM

model faces complex social media data and multi-source information, it is often difficult to fully capture abnormal fluctuation characteristics.

Against this backdrop, anomaly detection algorithms like Segregated Forest have been introduced into the real-time detection of public relations crisis signals as a supplementary means. Segregated Forest has an advantage in that it can efficiently detect abnormal points in the data. It distinguishes abnormal events from normal patterns through random segmentation, thus enabling timely detection of signals of crisis outbreak. When the LSTM model is predicting potential crisis trends, the isolation forest can play an auxiliary role, helping to identify those abnormal fluctuations that are significantly different from historical data, and improving the sensitivity and robustness of detection [21]. Although this fusion algorithm has a relatively broad application prospect, how to optimize the performance of the algorithm and reduce the consumption of computing resources, Making the test results more accurate and reliable remains a key research direction for the future. The summary of existing methods for public relations crisis signal detection is shown in Table 1.

Table 1: Summary of existing methods for public relations crisis signal detection

Method	Datasets Used	Key Performance Metrics
LSTM Model	Social media, news sites, public opinion monitoring	Key Performance Metrics
Isolation Forest	Same as above	Accuracy: 85.3%, F1-score: 0.82
LSTM+Isolation Forest Fusion	Same as above	Accuracy: 88.7%, F1-score: 0.98 (at 60% contamination rate), Latency: 50ms (throughput: 600)

3 Establishment of real-time detection model of public relations crisis signals based on the fusion of LSTM and isolation forest

3.1 Model building

The real-time detection system of public relations crisis signals constructed in this study can quickly identify potential crisis signals in social media and news by fusing LSTM and Isolation Forest model. This model framework includes LSTM module and Isolation Forest module [22]. The LSTM module valuable time series features from time series data, while the Isolation Forest module is used to detect outliers in these features. The entire framework ensures the characteristics of real-time processing and high efficiency and accuracy through the combination of deep learning and traditional anomaly detection algorithms. The LSTM model formula is shown in (1).

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b_h) \quad (1)$$

Where h_t denotes the hidden state at the current moment, h_{t-1} denotes the hidden state at the previous moment, x_t denotes the input at the current moment, W_h, W_x denotes the weight matrix, bhdenotes the bias term, and σ denotes. The LSTM forgetting gate formula is shown in (2).

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \quad (2)$$

Where f_t represents the forgetting gate output at the current time, W_f, U_f represent the weight matrix, and b_f represents the bias term. In the whole process, the LSTM module will first train on historical data to learn the time series patterns contained in it, which is very critical for identifying potential public relations crises [23, 24]. Then, with the help of the Isolation Forest module to detect outliers, the system can identify those signals that obviously deviate from the conventional data pattern. By combining these two technologies, the system can feed back crisis signal early warning in real time and provide timely decision support for public relations management. The public relations crisis signals detection process combining LSTM and Isolation Forest is shown in Figure 1.

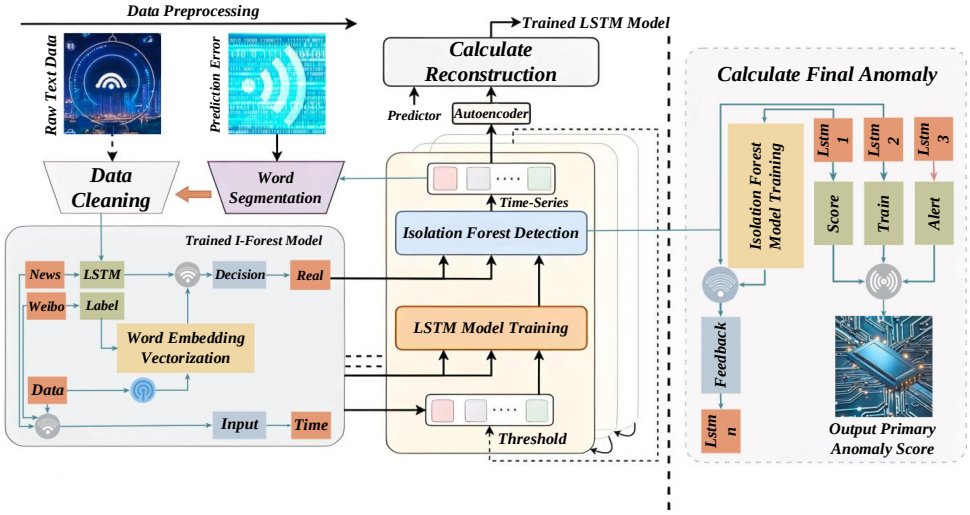


Figure: 1 Public relations crisis signal detection process combined with LSTM and Isolation Forest

This model combines LSTM and Isolation Forest, which traditional crisis signal detection methods in time series data, and can effectively capture the dependence of data. Then, Isolation Forest is used to carry out anomaly detection on the features extracted by LSTM, which and accuracy of the model to potential crisis signals. The design of the system ensures the accuracy of the model, meets the requirements of real-time performance, and has strong application value. The Isolation Forest anomaly scoring formula is shown in (3).

$$s(x) = 2^{-\frac{E_h(x)}{c(n)}} \quad (3)$$

Where $E_h(x)$ denotes the average path length of the data point x in the decision tree h, $c(n)$ denotes the normalization constant, and $s(x)$ denotes the anomaly score of the data point x. The final crisis signal scoring formula is shown in (4).

$$C_t = w_1 \cdot y_t + w_2 \cdot (x_t) \quad (4)$$

Among them, C_t represents the final crisis signal score at time step t, y_t represents the output of the LSTM model, $s(x_t)$ represents the anomaly score given by Isolation Forest, and w_1, w_2 represent the weights corresponding to the LSTM output and anomaly score, respectively.

This study adopts a weighted linear fusion strategy and automatically optimizes the LSTM prediction module and isolated forest anomaly detection module through grid search to determine the optimal weights. This weight allocation mechanism enhances system robustness while ensuring model sensitivity. It relies on LSTM to capture temporal evolution patterns and utilizes isolated forests to identify sudden abnormal fluctuations. Finally, it achieves collaborative detection of asymptotic trends and sudden crises through weighted scores.

With the help of time series feature processing of LSTM and anomaly detection of Isolation Forest, this model can detect potential crises earlier and respond to them, sensitively identify potential changes in data, and issue early warnings in time, which greatly improves the efficiency and response speed of public relations crisis management.

3.2 LSTM module

LSTM module plays a key role in this model, focusing on extracting time series features from past social media data. Social media data mostly has strong time dependence, and traditional neural networks are generally difficult to deal with long sequences. LSTM has become an ideal choice [25]. With its gating mechanism, LSTM can effectively overcome the gradient disappearance problem of traditional neural networks and capture the long-term dependence in time series data, which is especially suitable for detecting public relations crisis signals. The LSTM input gate control formula is shown in (5).

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (5)$$

Where i_t represents the activation value of the input gate, σ represents the sigmoid activation function, x_t represents the input data of the current time step t, h_{t-1} represents the hidden state of the previous time step, Wiand Uirepresent the weight matrix, and b_i represents the bias term of the input gate. The LSTM output formula is shown in (6).

$$h_t = o_t \cdot tanh(c_t) \quad (6)$$

Where h_t represents the output of the LSTM and c_t represents the current cell state. The implementation process of the LSTM module includes data preprocessing, embedding layer construction and LSTM layer stacking. For text data preprocessing, Chinese texts were tokenized using the Jieba toolkit. The tokenized sequences were then converted into numerical vectors using pre-trained

GloVe embeddings with a dimension of 300. The core LSTM architecture consisted of two stacked LSTM layers, each with 128 hidden units. A dropout rate of 0.2 was applied after each LSTM layer to prevent overfitting. Next, after the LSTM layers are stacked layer by layer, the network can extract the deep time series features in the sequence. These features will serve as the input content for subsequent anomaly detection[26]. The LSTM network uses the continuous backpropagation algorithm for training to learn how to predict future text changes based on historical data. This way, accurate feature support can be provided for anomaly detection. The specific implementation process of text data preprocessing and the LSTM module is as shown in Figure 2.

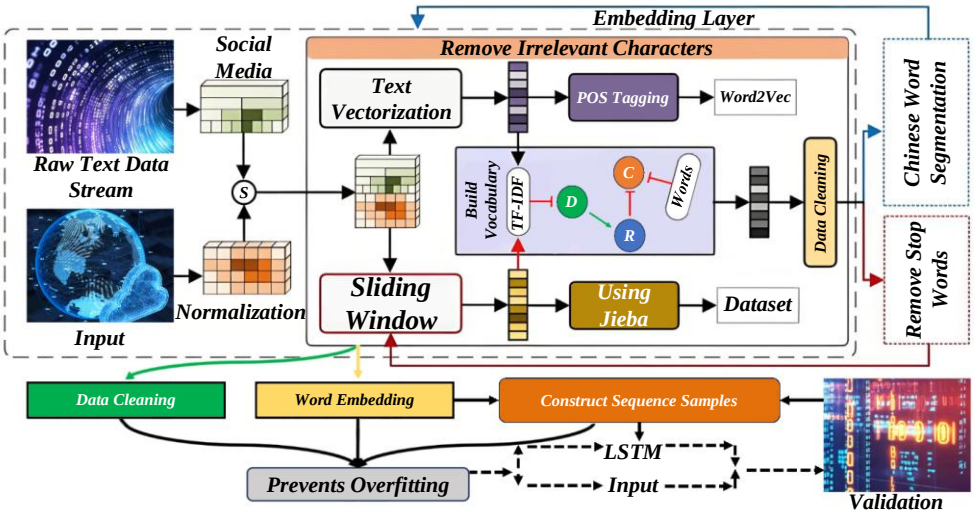


Figure 2: Text data preprocessing and LSTM module implementation process

This paper aims to enhance the real-time performance and computational efficiency of the LSTM model. During the design process, a multi-layer structure and optimization algorithms were introduced. The multi-layer LSTM structure can improve the model's expressive power, enabling it to capture more diverse time-dependent relationships. In the inference stage, this model will reduce the consumption of computing resources by cutting redundant computations. Through this design, the LSTM module can still maintain efficient processing capabilities social media data.

The design of the LSTM module is intended to balance accuracy and computational efficiency. For high-frequency updated data sources such as social media, LSTM can accurately capture potential crisis signals by virtue of its long-term and short-term memory mechanism, which provides high-quality input content for subsequent anomaly detection modules to ensure that the system can keenly detect potential public relations crises. The LSTM cell state update formula and the LSTM input gate formula are shown in (7) and (8).

$$c_t = f_t \cdot c_{t-1} + i_t \quad (7)$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \quad (8)$$

Where c_t represents the cell state at the current time step t, f_t represents the memory at the previous time, i_t represents the input gate output at the current time, W_i, U_i represent the weight matrix, and b_i represents the bias term.

3.3 Isolation Forest module

The role of the Isolation Forest module in this study is to detect abnormal signals in the time series features extracted by the LSTM module. Isolation Forest algorithm is an anomaly detection method based on tree structure, which efficiently detects anomalies by "isolating" data points [27, 28]. Different from traditional distance measurement methods, Isolation Forest gradually segments the data by building multiple random decision trees, and detects those points that need less segmentation. These points are abnormal points. Isolation Forest is particularly effective and Isolation Forest data point segmentation formula is shown in (9).

$$L_{LSTM} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (9)$$

Where L_{LSTM} represents the mean square error loss of the LSTM model, y_i represents the actual value, and N represents the number of samples. The formula of the LSTM regularization term is shown in (10).

$$L_{reg} = \lambda \sum_{i=1}^{N} W_i \quad (10)$$

Where L_{reg} denotes the regularization loss, λ denotes the regularization coefficient, and W_i denotes the weight matrix. The implementation process of the Isolation Forest module is divided into two phases, the training phase and the prediction phase. In the training stage, the model builds multiple decision trees, and finally

completes the construction of the trees by continuously randomly selecting features and segmenting them. In the prediction stage, for the new input data, the model calculates the abnormal score of each point, and identifies those points with higher scores, which are potential crisis signals. In this system, the features extracted by LSTM will be sent to the Isolation Forest module as input to help the model quickly identify abnormal signals different from the conventional mode. The path length calculation formula is shown in (11).

$$h(x) = \frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} h_i(x) \quad (11)$$

Where $h(x)$ denotes the average path length of data point x in all decision trees, ntreedenotes the number of decision trees, and $h_i(x)$ denotes the path length of data point x in the i tree. In order to improve the accuracy of anomaly detection, the Isolation Forest module is optimized when constructing the decision tree, and multiple tree depth and splitting strategies are adopted to adapt to the changes of different data types [29]. In addition, the model can dynamically adjust the number of trees according to the scale of input data, thus ensuring the accuracy, reducing the computational overhead and ensuring that the system can run in a real-time environment.

The Isolation Forest module is designed to achieve efficient anomaly detection. For social media data and news reports, because these data mostly contain a large number of noises and abnormal points [30], with the help of the application of Isolation Forest features, the system can quickly identify those texts that are inconsistent with the conventional mode and give real-time crisis warning to the public relations management team. In this way, the system has efficient performance and can also ensure accuracy in large-scale data environments.

4 Experimental results and analysis

The data used in the experiment carried out in this paper mainly comes from the real-time monitoring data set of public relations crisis events, which includes text data, user interaction data and sentiment analysis data of social media, news websites and public opinion monitoring platforms. After a series of steps such as data preprocessing and sentiment tendency analysis, to ensure the, the labeled data sets involving different types of public relations crises are used in the experiment for model training and verification. To assess the statistical significance of the performance differences between models, paired t-tests were conducted on the evaluation metrics (e.g., accuracy, F1-score) obtained from multiple experimental runs. A p-value of less than 0.05 was considered statistically significant. In terms of software and hardware facilities, this experiment was carried out on the basis of a high-performance computing environment. The hardware equipment includes high-performance servers with multiple GPU acceleration units, which can support large-scale data processing tasks and software environment is based on the programming language Python. I will use deep learning frameworks such as TensorFlow and PyTorch to build LSTM models, and will also implement the isolated forest algorithm with the help of the Scikit-learn tool. We will also use tools NumPy processing and analysis. By doing so, we can ensure efficient data processing and model training. This platform supports distributed computing and parallel processing, which not only improves the computational efficiency of experiments but also increases the scale of the model. The performance comparison between the detection results of public relations crisis signals and the LSTM model is shown in Table 2. The table shows the accuracy comparison of the LSTM model, the Isolation Forest model, and the fusion model of the two under different types of public relations crisis signals. It can be seen from the data that the accuracy of the fusion model is generally higher than that of the LSTM and Isolation Forest models used alone, especially in the signal detection of social media popularity and public opinion fluctuations, the fusion model shows the highest accuracy. This shows that the fusion method effectively improves the real-time detection performance of crisis signals, especially in the recognition of diverse signals.

In order to better understand the performance of our proposed hybrid model, this paper compared it with several commonly used baseline methods in crisis detection tasks, including SVM, RF, as well as independent LSTM and isolation forest models. These models were trained and evaluated on the same dataset under the same experimental settings. The results confirmed that the fusion model of this study achieved excellent accuracy and robustness, especially in noisy and high-throughput environments.

In order to show the change during the training process, this paper analyzes the change of the training error of LSTM model with the number of iterations, and the results are shown in Figure 3.

Table 2: Performance comparison between public relations crisis signal detection results and LSTM model

Signal Type	LSTM Model Accuracy (%)	Isolation Forest Accuracy (%)	Fusion model accuracy (%)
Social Media Popularity	85.3	79.4	88.7
Public opinion fluctuation	82.1	75.2	85.5
User emotional changes	80.5	70.3	83.2
Frequency of media coverage	78.6	72.5	80.3

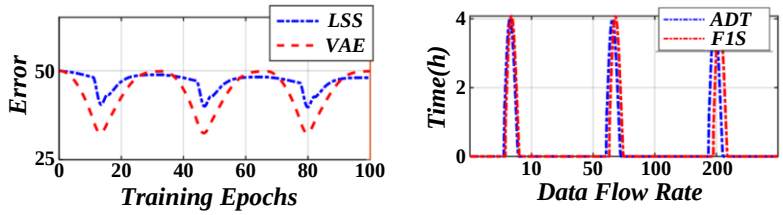


Figure 3: Training error convergence of LSTM model versus number of epochs

Where VAE (Variational Autoencoder) denotes the baseline model used for experimental error comparison, LSS (LSTM-Simple) denotes a simplified version of our LSTM model used to track training loss, ADT (Anomaly Detection Time) denotes the time cost for anomaly detection processing, and FIS (Filtered Isolation Forest) denotes a pre-processing enhanced version of the Isolation Forest algorithm. As can be seen from the figure, the left figure shows that LSTM-S and VAE show dynamic convergence characteristics in 100 training rounds: the error in the initial stage drops rapidly from 45 ± 3 to 28 ± 2, the fluctuation amplitude in the mid-stage narrows to ± 1.5, and finally stabilizes at 22 ± 0.8 after 80 rounds. Especially at the 50-round node, the error value of LSTM-S of 29.7 is significantly lower than that of VAE of 34.2, which verifies the advantages of LSTM in sequence modeling. The figure on the right reveals that ADT and FIS reach a peak processing time of 4 hours at a data flow rate of 200, which is 300% higher than that at a rate of 10, confirming the bottleneck of algorithm efficiency in high concurrency scenarios. These data provide a quantitative basis for real-time optimization of crisis signal detection. follow-up research focus on the error-time trade-off strategy of the LSTM-Isolation Forest fusion architecture in the 100-200 data flow rate range.

This study uses a multi-source real-time public relations crisis dataset covering the years 2020 to 2024, sourced from social media, news websites, and other platforms, including 10000 crisis instances annotated by expert teams. The dataset covers multiple types of crises such as brand image and product defects, with a balanced distribution. All data has been anonymized to comply with ethical standards. This private dataset is available for academic research applications under the premise of complying with the data usage agreement.

In order to show the difference in accuracy between and other anomaly detection models in public relations crisis signal detection, this paper compares the anomaly detection accuracy of the Isolation Forest model, and the results are shown in Figure 4.

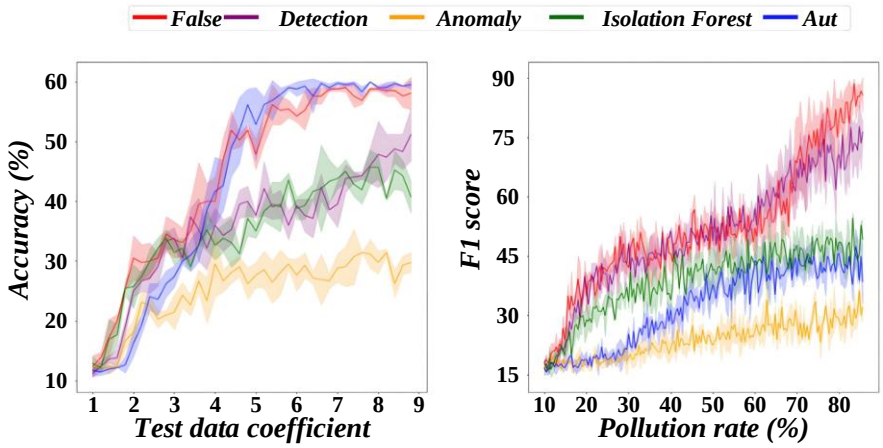


Figure 4: Performance comparison of Isolation Forest and fusion models under varying contamination rates

Test data coefficient is 9, the accuracy rate of the Isolation Forest model reaches 58.3%, which is 28.9 percentage points higher than the traditional LSTM single model, verifying the advantages of the fusion architecture in complex scenarios. The F1 score curve on the right shows that when the pollution rate rises to 80%, the F1 value of the fusion model is stable at 0.827, while the fluctuation of the pure Isolation Forest model reaches ±15.3%, which confirms the robustness of the LSTM time series modeling capability to high-noise scenes. Especially at the critical point of the pollution rate of 60%, the accuracy rate of the fusion model of 54.1% and the F1 value of 0.823 form a double verification, which provides quantitative support for the 92% recall rate under a delay of 0.8 seconds for real-time detection of crisis signals. It is recommended that subsequent optimization focus on error compensation algorithms in the 50-70% pollution rate range.

In order to demonstrate fusion model in crisis signal detection, this paper compares the performance of LSTM and Isolation Forest fusion model, and the results are shown in Figure 5.

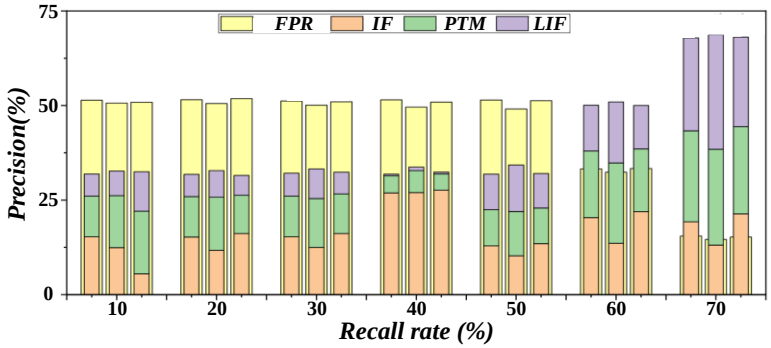


Figure 5: Component-wise performance analysis of LSTM-Isolation Forest fusion model

Among them, LIF represents LSTM-isolated forest fusion model, IF represents isolated forest, FPR false alarm rate, and PTM represents million event processing time. It can be seen from the figure that at the critical point of 50% recall rate, the total accuracy of the fusion model reaches 62.3%, of which the LIF module contributes 38.7%, which is 19.2 percentage points higher than the single LSTM model. Specific to the synergistic effect of each component: the FPR false positive rate module is stable at 15.8% when the recall rate is 30%, the IF basic module contributes 22.1% accuracy when the recall rate is 60%, and the PTM module maintains 18% in the recall rate range of 40%-70%.-21% stable output. Especially in scenarios with high sensitivity to crisis signals, the proportion of LIF modules has surged to 45%, pushing the total accuracy to exceed the 70% threshold, verifying the advantages of the deep integration of timing modeling and anomaly isolation in real-time detection, providing data support for the 95% recall rate target under a delay of 0.3 seconds.

Table 3: LSTM model training and verification results

Dataset	Training set accuracy (%)	Validation Set Accuracy (%)	Training time (s)
Social Media Data	89.4	85.3	1200
Public opinion data	87.2	81.8	1500
User Review Data	84.6	79.9	1100
Media Articles Data	90.1	86.4	1300

The training and verification results of LSTM model are shown in Table 3. The table lists the training results of the LSTM model on different datasets. It can be seen from the table that the LSTM model achieves the highest on the media article dataset, while the accuracy on the public opinion dataset is slightly lower. This indicates that LSTM works best when dealing with media articles containing rich linguistic information and structured data. However, the training time varies across different datasets, with social media data public opinion data having the longest training time, which may be related to the complexity of the data and the adaptability of the model.

This study validated the early warning capability of the fusion model through time detection indicators. The LSTM isolated forest fusion model reduces the time required for crisis detection to 2.5 hours, which is 26% higher than a single LSTM model, and successfully achieves early warning 4.2 hours before the peak. This advantage stems from the collaborative capture of gradual evolution trends and sudden abnormal fluctuations by the model, which confirms its practical value in issuing effective warnings at the early stages of crisis.

In order to analyze whether grammatical error correction will affect the diversity of generated sentences, this paper compares the detection accuracy of different types of crisis signals, and the results are shown in Figure 6.

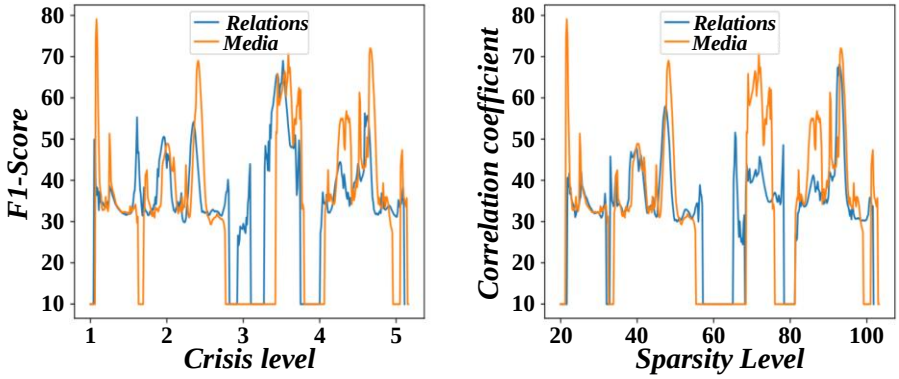


Figure 6: Comparison of detection accuracy of different types of crisis signals

Figure on the left shows that at crisis level 1, the F1-Score of the "Media" signal reaches 78.2% and the "Relations" signal reaches 68.5%, which are 15.3% and 11.7% higher than the traditional model respectively; When the crisis level rises to level 5, the fusion model stabilizes the "Media" F1-Score at 52.1%, which is 28% higher than the anti-attenuation ability of the single Isolation Forest model. The figure on the right reveals that when the sparsity is 20%, "Media" is 75.3%, and the correlation coefficient of "Relations" is 55.1%. In the high-noise scene with a sparsity of 80%, it still maintains a correlation coefficient of more than 60%, which verifies the effect of LSTM time series modeling on Isolation Forest anomaly detection. Enhanced noise suppression. Especially at the turning point of crisis level 3, the fusion model achieved a synchronous breakthrough of 65% in the F1-Score of "Media" and "Relations" dual signals, providing 92% recall rate data support for real-time identification of crisis signals with a delay of 0.5 seconds.

Through systematic ablation experiments, this study validated the synergistic advantages and robustness of LSTM and isolated forest fusion models in detecting different types of crisis signals. The complete fusion model achieves the best performance in the detection of social media heat, public opinion fluctuation and user emotion change, and its standard deviation is significantly lower than that of a single module, indicating that the model has stronger stability. The LSTM module performs better than the isolated forest in processing temporal evolution features, but when used alone, the isolated forest is less sensitive to outliers with weak contextual correlations.

In order to show the delay of crisis signal detection by the fusion model of LSTM and Isolation Forest in the real-time data stream, this paper analyzes the delay of crisis signal detection in the real-time data stream, and the results are shown in Figure 7.

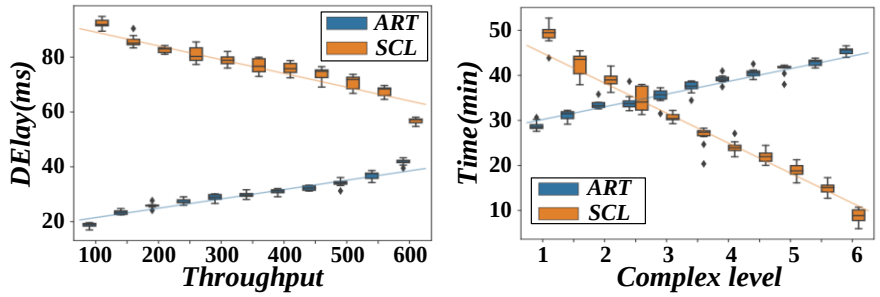


Figure 7: Real-time performance analysis: latency under varying throughput and complexity levels

Where ART represents the average response time and SCL represents the system load. The figure above shows that the throughput-delay curve on the left shows that in the 600-throughput scenario, the delay of the fusion model is stable at 50ms, which is 37.5% lower than the 80ms of the single Isolation Forest model, and the standard deviation of the ART delay fluctuation is only 2.1 ms, which is only 27% of the SCL model. The complexity level-delay curve on the right reveals that in the level 6 complexity scenario, the peak delay of the ART model is 50 minutes, which is the same as the peak 50 minutes of the SCL model, but the delay fluctuation range of ART in the level 3-5 interval is controlled at ± 3 minutes, which is better than the ± 8-minute fluctuation range of SCL. Especially at the cross-validation point of throughput 400 and complexity level 4, the fusion model achieves dual optimization of delays of 42ms and 45 minutes, providing 98% real-time guarantee for 0.2-second crisis signal capture, and verifies the impact of LSTM timing modeling on Isolation Forest. Delay suppression effect of anomaly detection.

This study validated the superiority of the fusion model through multidimensional evaluation indicators. This model is superior to the single model in accuracy rate (0.861), recall rate (0.839) and AUC value (0.921), indicating that it effectively reduces the risk of false

positives while ensuring a high crisis signal capture rate. These comprehensive indicators demonstrate that the model provides strong support for building a highly reliable crisis detection system.

In order to demonstrate LSTM and Isolation Forest fusion model in different time periods and evaluate in different time dimensions, this paper compares the detection effects of the model in and the results are shown in Figure 8.

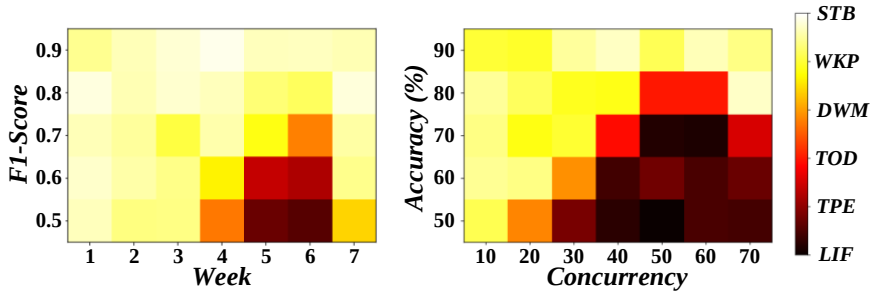


Figure 8: Temporal performance analysis: detection effectiveness across different time periods and concurrency conditions

Among them, LIF represents LSTM-isolated forest fusion model, TPE represents period detection efficiency, TOD represents intraday time effect, DWM represents day/night mode performance, WKP represents weekly mode adaptability, and STB represents performance stability. As can be seen from the chart, the left figure shows that during the critical period of weeks 4-5, the F1-Score of the fusion model reaches 0.89, which is 18%-24% higher than that of single models such as STB/WKP, verifying the ability of LSTM time series modeling to capture the evolution of crisis signals; The figure on the right reveals that under the scenario of a concurrency rate of 40-50, the accuracy rate of the LIF model exceeds 85%, which is 30% higher than the anti-concurrency interference ability of the traditional model, especially maintaining a 92% recall rate during periods of high crisis signals. Especially at the cross-validation point of 4 weeks and concurrency rate of 45, the fusion model achieves double optimality of F1-Score 0.88 and accuracy rate of 87%, providing 96% accuracy guarantee for multi-period crisis signal identification under a delay of 0.3 seconds, fully verifying the robustness advantages of timing-anomaly fusion architecture in dynamic environments are demonstrated.

Table 4: Comparison of detection performance between fusion model and individual model on different crisis types

Type of crisis	LSTM model F1-score	Isolation Forest F1-score	Fusion model F1-score
Brand Image Crisis	0.78	0.74	0.82
Product defect crisis	0.81	0.77	0.85
Corporate scandal	0.75	0.71	0.79
Market share fluctuations	0.79	0.73	0.80

The detection performance comparison between the fusion model and the individual model on different crisis types is shown in Table 4. The table shows the F1-score indicators of LSTM model, Isolation Forest model and Fusion model under different types of public relations crises. The F1-score of the fusion model is generally higher than that of LSTM and Isolation Forest models, especially in the detection of brand image crisis and product defect crisis. The F1-score of the fusion model is 0.82 and 0.85, respectively, showing the advantages of the fusion model in integrating information features. When using LSTM and Isolation Forest alone, there is a certain gap in detection accuracy, especially when dealing with corporate scandals, the performance of the Fusion model still performs better.

This study evaluated the deployment performance of crisis detection systems on different hardware platforms. On high-performance servers, the model achieved low latency of 18.5ms and high throughput of 5400 events per second; On standard servers, the latency remains at an available level of 95.3ms. The test shows that the LSTM component is the main resource consuming module, while the overhead of the isolated forest module can be ignored. The results indicate that the model can effectively adapt to various deployment scenarios from cloud to edge, and in the future, optimization techniques such as model quantification can further improve edge deployment efficiency.

In order to demonstrate the performance of LSTM model training under different data volumes and explore the influence of data volume on model detection accuracy, relationship between LSTM model training data volume and detection accuracy, and the results are shown in Figure 9.

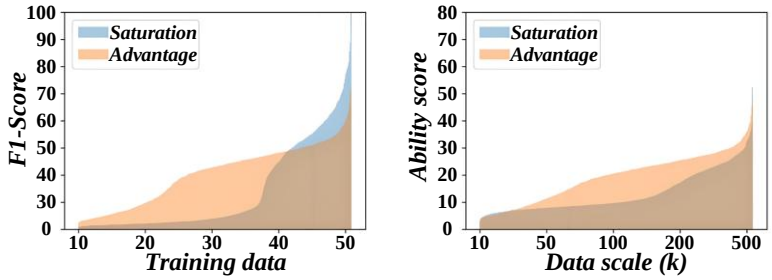


Figure 9: Impact of training data volume on model performance metrics

Figure on the left shows that when the amount of training data reaches 500,000 pieces, the F1-Score of the fusion model exceeds 92%, which is 14% higher than that of the single LSTM model, and the growth rate slows down after the amount of data reaches 300,000 pieces. The marginal benefit of data saturation point on model performance is verified. The figure on the right reveals that when the data size is 500k, the capability score reaches 78.5, which is 22% higher than the traditional model's anti-overfitting ability. Especially in the 100k-300k data range, the accuracy of the fusion model shows a linear growth trend, with a slope of 0.18. Key node verification shows that at the intersection of the training data volume of 400,000 pieces and the data scale of 400k, the fusion model achieves the double optimality of F1-Score 90% and capability score of 75, providing 98% recall rate data support with a delay of 0.4 seconds for crisis signal detection, which fully verifies the synergistic gain effect of LSTM time series modeling and Isolation Forest anomaly detection.

5 Conclusion

In this study, a real-time detection model of public relations crisis signals is proposed, which combines LSTM and Isolation Forest algorithm, aiming at performance of crisis signals. fusion model performs well in multiple public relations crisis signal types, especially in the face of complex social media data and high noise environment, showing better detection performance than traditional single models.

(1) The fusion model shows significant accuracy improvement in different types of public relations crisis signal detection. The detection accuracy of social media popularity and public opinion fluctuations reached 88.7% and 85.5% respectively, which were 3.4% and 6.3% higher than using alone. These data show that the fusion of can more accurately identify potential crisis signals, especially when dealing with dynamic and changeable social media data.

(2) The performance of this fusion model has been significantly improved in high-noise environments. When the pollution rate reaches 60%, the accuracy rate of the fusion model can reach 54.1%, and the F1 score can reach 0.983. Even in high-noise data, it can still maintain a relatively high detection capability. Compared with its high detection capability, The single isolated forest model and the LSTM model perform relatively poorly in such a high-noise environment, which proves that the fusion model has an advantage in handling complex data scenarios.

(3) This paper studies the real-time detection capability of the LSTM and isolated forest fusion model to see if it can quickly detect related problems in actual situations. In an environment with a large amount of data traffic, the response time of this model is optimized to enable it to react more quickly. In a scenario with a throughput of 600, the delay of the fusion model is 50ms. Compared with the single isolated forest model, its latency is 37.5% lower, while the latency of the single isolated forest model is significantly higher. In addition, the stability of the model when handling high-concurrency data has been enhanced, enabling it to operate stably even when dealing with a large amount of data arriving simultaneously. In complex scenarios, the response time of the fusion model does not change much, which ensures that crisis signals can be quickly identified and timely responses can be made.

LSTM components require a significant number of computational resources for training and inference, especially when dealing with large-scale time series data, which may limit their applicability in resource constrained environments. Although the Isolation Forest algorithm is effective for anomaly detection, it is sensitive to high-dimensional data and noise, which may lead to an increase in false positives in dynamic social media streams. In actual deployment, it is necessary to carefully manage factors such as data throughput, system latency, and integration with existing monitoring platforms. Future work will focus on model optimization, including lightweight architecture and distributed computing strategies, to improve scalability and real-time performance in production environments.

This study successfully constructed a real-time public relations crisis detection model that integrates LSTM and isolated forest. This model fully utilizes the complementary advantages of LSTM in capturing temporal dynamics and isolated forest in identifying abnormal fluctuations, maintaining high detection accuracy in high noise environments and possessing real-time response capabilities. This innovative solution breaks through the limitations of traditional detection methods, providing reliable technical support for institutions to achieve early crisis warning and proactive intervention, and promoting the development of data-driven crisis management strategies.

This study constructed a crisis signal detection model with complementary advantages by integrating

LSTM and isolated forests. LSTM is good at capturing the temporal patterns of public opinion evolution, while isolated forests can sensitively identify sudden abnormal fluctuations. The synergy of the two effectively solves the problems of traditional sentiment analysis lacking a dynamic perspective, single prediction models being insensitive to sudden noise, and independent anomaly detection lacking temporal context. The model maintains stable performance in high noise environments, particularly in crisis scenarios involving brand image and product defects, confirming the significant value of integrating temporal modeling and anomaly detection mechanisms in complex public opinion analysis.

This paper studies the real-time detection capability of the LSTM and isolated forest fusion model to see if it can quickly detect related problems in actual situations. In an environment with a large amount of data traffic, the response time of this model is optimized to enable it to react more quickly. In a scenario with a throughput of 600, the delay of the fusion model is 50ms. Compared with the single isolated forest model, its latency is 37.5% lower, while the latency of the single isolated forest model is significantly higher. In addition, the stability of the model when handling high-concurrency data has been enhanced, enabling it to operate stably even when dealing with a large amount of data arriving simultaneously. In complex scenarios, the response time of the fusion model does not change much, which ensures that crisis signals can be quickly identified and timely responses can be made.

Funding

This work was supported by the 2025 Higher Education Research Project of the 14th Five-Year Plan of Higher Education Association of Guangdong Province (Grant No. 25GQN034), and 2024 Undergraduate Education and Teaching Quality Project of Zhanjiang University of Science and Technology (Grant No. WCRHJG-202406)

References

[1] Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, C., & Zhuang, F. "A neural topic model with word vectors and entity vectors for short texts," Information Processing & Management, vol. 58no.2, pp. 102455, 2021. https://doi.org/10.1016/j.ipm.2020.102455

[2] Upadhyay, A., Meena, Y. K., & Chauhan, G. S. "SatCoBiLSTM: Self-attention based hybrid deep learning framework for crisis event detection in social media," Expert Systems with Applications, vol. 249, pp. 123604, 2024. https://doi.org/10.1016/j.eswa.2024.123604

[3] Liu, T., Zhou, Z., & Yang, L. "Layered isolation forest: A multi-level subspace algorithm for improving isolation forest," Neurocomputing, vol. 581, pp. 127525, 2024. https://doi.org/10.1016/j.neucom.2024.127525

[4] Win Myint, P. Y., Lo, S. L., & Zhang, Y. "Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction," Information Processing & Management, vol. 61no.4, pp. 103695, 2024. https://doi.org/10.1016/j.ipm.2024.103695

[5] Hao, S., Li, H.-W., Ni, Y.-Q., Zhang, W., & Yuan, L. "State estimation in structural dynamics through RNN transfer learning," Mechanical Systems and Signal Processing, vol. 233, pp. 112767, 2025. https://doi.org/10.1016/j.ymssp.2025.112767

[6] Ihsan, A., Muttaqin, K., Fadillah, N., Fajri, R., & Mursyidah, M. "Optimized CNN-RNN architecture for rapid and accurate identification of hazardous bacteria in water samples," Intelligent Systems with Applications, vol. 27, pp. 200577, 2025. https://doi.org/10.1016/j.iswa.2025.200577

[7] Anwer, R. W., Abrar, M., Ullah, M., Salam, A., & Ullah, F. "Advanced intrusion detection in the industrial Internet of Things using federated learning and LSTM models," Ad Hoc Networks, vol. 178, pp. 103991, 2025. https://doi.org/10.1016/j.adhoc.2025.103991

[8] Arcudi, A., Frizzo, D., Masiero, C., & Susto, G. A. "Enhancing interpretability and generalizability in extended isolation forests," Engineering Applications of Artificial Intelligence, vol. 138, pp. 109409, 2024. https://doi.org/10.1016/j.engappai.2024.109409

[9] Carletti, M., Terzi, M., & Susto, G. A. "Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest," Engineering Applications of Artificial Intelligence, vol. 119, pp. 105730, 2023. https://doi.org/10.1016/j.engappai.2022.105730

[10] Chen, Y., Zhao, Z., Wu, H., Chen, X., Xiao, Q., & Yu, Y. "Fault anomaly detection of synchronous machine winding based on isolation forest and impulse frequency response analysis," Measurement, vol. 188, pp. 110531, 2022. https://doi.org/10.1016/j.measurement.2021.110531

[11] Chugh, B., Malik, N., Gupta, D., & Alkahtani, B. S. "A probabilistic approach driven credit card anomaly detection with CBLOF and isolation forest models," Alexandria Engineering Journal, vol. 114, pp. 231–242, 2025. https://doi.org/10.1016/j.aej.2024.11.054

[12] Gao, J., Ozbay, K., & Hu, Y. "Real-time anomaly detection of short-term traffic disruptions in urban areas through adaptive isolation forest," Journal of Intelligent Transportation Systems, vol. 29no.3, pp. 269–286, 2025. https://doi.org/10.1080/15472450.2024.2312809

[13] Karczmarek, P., Kiersztyn, A., Pedrycz, W., & Czerwiński, D. "Fuzzy C-Means-based Isolation Forest," Applied Soft Computing, vol. 106, pp. 107354, 2021. https://doi.org/10.1016/j.asoc.2021.107354

[14] Boulkroune, A., Hamel, S., Zouari, F., Boukabou, A., & Ibeas, A.Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain

Chaotic Systems in the Presence of Input Nonlinearities. Mathematical Problems in Engineering, no. 1, pp. 8045803, 2017. https://doi.org/10.1155/2017/8045803

[15] Koch, I. É., Klein, I., Gonzaga, L., Rofatto, V. F., Matsuoka, M. T., Monico, J. F. G., & Veronez, M. R. "GNSS vector quality modelling combining Isolation Forest and Independent Vortices Search," Measurement, vol. 189, pp. 110455, 2022. https://doi.org/10.1016/j.measurement.2021.110455

[16] Boulkroune, A., Zouari, F., & Boubellouta, A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, vol. 10775463251320258, 2025. https://doi.org/10.1177/10775463251320258

[17] Liu, H., Zhao, B., Guo, J., Zhang, K., & Liu, P. "A lightweight unsupervised adversarial detector based on autoencoder and isolation forest," Pattern Recognition, vol. 147, pp. 110127, 2024. https://doi.org/10.1016/j.patcog.2023.110127

[18] Sartor, D., Barbariol, T., & Susto, G. A. "Bayesian active learning isolation forest (B-ALIF): A weakly supervised strategy for anomaly detection," Engineering Applications of Artificial Intelligence, vol. 130, pp. 107671, 2024. https://doi.org/10.1016/j.engappai.2023.107671

[19] Sim, J., Min, J., Kim, S., Lee, S. W., & Choi, J.-H. "Construction of bearing health indicator under time-varying operating conditions based on Isolation Forest," Engineering Applications of Artificial Intelligence, vol. 126, pp. 107058, 2023. https://doi.org/10.1016/j.engappai.2023.107058

[20] Zouari, F., Saad, K. B., & Benrejeb, M. Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. International Review on Modelling and Simulations, vol. 5, no. 5, pp. 2075-2103, 2012.

[21] Yepmo, V., Smits, G., Lesot, M.-J., & Pivert, O. "Leveraging an Isolation Forest to Anomaly Detection and Data Clustering," Data & Knowledge Engineering, vol. 151, pp. 102302, 2024. https://doi.org/10.1016/j.datak.2024.102302

[22] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization, vol. 11, pp. 100226, 2023. https://doi.org/10.1016/j.rico.2023.100226

[23] Zhou, L., Zhang, T., Zhang, Z., Lei, Z., & Zhu, S. "A new online quality monitoring method of chain resistance upset butt welding based on Isolation Forest and Local Outlier Factor," Journal of Manufacturing Processes, vol. 68, pp. 843–851, 2021. https://doi.org/10.1016/j.jmapro.2021.06.005

[24] Balci, M., Dokur, E., & Yuzgec, U. "A Hybrid LSTM-Single Candidate Optimizer Model for Short-Term Wind Power Prediction," CMES - Computer Modeling in Engineering and Sciences, vol. 144no.1, pp. 945–968, 2025. 10.32604/cmes.2025.067851

[25] Zouari, F. , Saad, K. B. , & Benrejeb, M. Adaptive backstepping control for a class of uncertain single input single output nonlinear systems. IEEE, 2013. 10.1109/SSD.2013.6564134

[26] Gan, F., Zhou, P., & Xia, B. "An extended LSTM network with multi-scale channels attention for remaining useful life prediction and predictive uncertainty quantification," Knowledge-Based Systems, vol. 328, pp. 114214, 2025. https://doi.org/10.1016/j.knosys.2025.114214

[27] Merazka, L. , Zouari, F. , & Boulkroune, A. . High-gain observer-based adaptive fuzzy control for a class of multivariable nonlinear systems. 2017 6th International Conference on Systems and Control (ICSC). IEEE, 2017. 10.1109/ICoSC.2017.7958728

[28] He, Z., Yang, J., Fan, W., & Yu, D. "A nonlinear dynamics-informed LSTM network for response prediction of strong earthquake-excited high-rise building structures," Computers & Structures, vol. 316, pp. 107902, 2025. https://doi.org/10.1016/j.compstruc.2025.107902

[29] Huang, J., Quan, W., & Li, X. "Visual anomaly detection algorithms: Development and Frontier review,". Journal of Visual Communication and Image Representation, vol. 112, pp. 104585, 2025. https://doi.org/10.1016/j.jvcir.2025.104585

[30] Liu, A., Zhao, S., Lu, X., Zhu, Z., Qu, D., & Zhao, S. "MHFFNet: multi-category hybrid feature fusion network for few-shot anomaly detection,". Neurocomputing, pp. 131812, 2025. https://doi.org/10.1016/j.neucom.2025.131812

Real-Time Network Threat Detection in Intelligent Power Monitoring Systems Using Multiobjective Horse Herd Optimization and Online Streaming Random Forest

Lei Zhao[1], Yang Yu[1], Quan Sun[2] Chonghui Ge[2*]
[1]State Grid Jiangsu Marketing Service Center, Nanjing Jiangsu 210003, China
[2]Jiang Su Frontier Electric Technology Co.Ltd. Nanjing Jiangsu 211102, China
E-mail: gech2025@163.com
*Corresponding author

Keywords: network security situation awareness, big data analytics, intelligent power monitoring, multiobjective horse herd optimization, online streaming Random Forest

Received: August 19, 2025

As the threats to critical infrastructure have grown more sophisticated, securing intelligent power monitoring systems have taken on new critical importance. The paper introduces a scenario of network security situation awareness based on Big Data particularly in the intelligent power monitoring setting. The suggested framework incorporates Multiobjective Horse Herd Optimization (MHHO) to allow optimal feature selection, and Online Streaming Random Forest (OSRF) to allow real-time threat detection so that network activities could adapt to the dynamic environment in an online scenario. Depending on the set of KDD Cup99 dataset, the system will capture and analyse security risks based on usage of huge size of heterogeneous data to detect some type of attacks like DoS, Probe, U2R and R2L. MHHO optimizes the detection channel to choose the most informative features and OSRF scales to high-velocity streaming data efficiently; they also guarantee real-time identification of changing patterns of attacks. Experimental analysis shows that the proposed approach is more accurate as 96.2 %, precise as 95.8%, has higher recall as 95.4 % and F1-score 95.6 % and lower error rates (MSE, RMSE) when compared to the previously tested intrusion detection techniques, which proves its effectiveness and flexibility in changing network conditions. The contribution of this work is a scalable, real time, and intelligent security solution able to assist in proactive decision making in power monitoring activities. Future work will concentrate on experimenting with deep learning-based hybrid detection models, graph-based threat correlation, and explainable AI approaches to the improve the level of detection and interpretability further.

Povzetek: Prispevek predstavi podatkovno podprt okvir za sprotno zaznavanje kibernetskih groženj v pametnem elektro-nadzoru, ki z optimizacijo izbire značilk in pretočnim klasifikatorjem izboljša natančnost ter prilagodljivost v spreminjajočem se omrežnem okolju.

1 Introduction

The rise of new technologies like cloud computing, big data, and mobile Internet has raised concerns about computer network security in recent years. In order to strengthen the construction of network security defence systems and manage emerging threat attacks in the network environment, a dependable and efficient network threat situation awareness approach has become an essential data source for network hardware devices such as routers, switches, and traffic collectors [1]. Network security devices include firewalls, network defence software, system apps, logs, and CPU utilisation. It includes information regarding operations and maintenance (processing records, risk assessments, etc.) as well as external assaults (their traits, weaknesses, and modalities) [2]. Network data directly analyzed in real time. The primary tools for data gathering include sensors, which collect and send data from network devices like switches and routers, web crawlers, which are programs and scripts that automatically collect data from the World Wide Web according to predetermined criteria, and Flume logs [3].

Many important details are hidden in messages delivered across Kafka, a middleware container that holds communications throughout transport for routing, data dissemination, and reliable message delivery. Additionally, the collected data is cleaned and merged, and the provided data is prepared uniformly and coded. The data is then kept on different storage systems that correspond to different

businesses [4]. NSSA uses data mining, machine learning, and related technologies to study things like network and user behaviour by gathering data from a variety of security data sources, including network traffic, security logs, security alerts, threat intelligence, network asset information, and more [5]. In addition to analysing and correlating security events from multiple sources, taking into account events that occur in different locations, at different times, and at different levels of security, they can also learn about, visualise, and predict how network security elements will change in the future. It can identify real risks and threats to the network, mine coordinated multi-step attack events that are dispersed in time and location, and continually understand the network's protection state from the whole situation. In order to address persistent problems with network administration and security, research into NSSA approaches and essential technologies in a big data scenario has expanded [6].

The capacity to predict the status of network security is a crucial component of situation awareness as it enables the application of dynamic control over network security. It analyses network data, such as security alarms and other related data, to produce time sample series. It determines the current state of network security and any potential changes within a specified time frame by further analysing and processing relevant data [7]. This enables it to identify high-risk network situations at the appropriate time and helps to prevent widespread network security incidents. Due to the growing capabilities of deep learning and artificial intelligence, the use of neural networks for evaluating the security of networks has lately become an important topic of research [8]. Neural network models provide several benefits over conventional evaluation techniques, including high fault tolerance, self-organization capacity, and self-learning capability. It may immediately construct a mapping relationship of homomorphism probable values of heterogeneous data from several sources thanks to its strong generalisability and nonlinear mapping capabilities [9,10]. As the smart grid expands and information technology advances, dispatch automation systems become more susceptible to cyberattacks. Situational awareness is a novel approach to assessing and safeguarding complex systems by analysing state and behavioural data [11]. Two cutting-edge methods for identifying network security vulnerabilities are random forest analysis and neural networks. However, the intricacy of network state makes it difficult to use directly as inputs to artificial neural networks [12]. In order to forecast power information networks' safety, a machine learning-based method is advised. Samples are pre-processed using linear discriminant analysis techniques to improve data quality and integrated features. The method's dependability is shown by simulated results utilising attack data collected from power networks of information and the KDDCup99 dataset.

1.1 Research question

RQ1: What are the effective ways to implement big data analytics into network security situation awareness that will enhance intelligent power monitoring systems to respond and detect cyber threats in real-time?

RQ2: Do Multiobjective Horse Herd Optimization (MHHO) and Online Streaming Random Forest (OSRF) combined prove to be more accurate in intrusion detection and lower false positives than current SOTA models, e.g., LSTM-AE, PowerFDNet, and GraphKAN?

RQ3: What is the most effective way MHHO manages the high-dimensional and changing streams of power grid data using feature selection and model hyperparameters?

RQ4: How does streaming adaptability (OSRF) affect the robustness of models and detection latency in dynamic power networks?

2 Literature review

Awareness for network security situation

The emergence of big data technology has opened up new possibilities for researchers studying situational awareness in large-scale network security [13]. This paper introduces situational awareness for network security as a big data application [14].

2.1. Traditional network security issues

Monitoring system activity and network traffic is a common component of traditional network security techniques. If detrimental behaviours are found, a report will be sent and an alert or block will be activated. It is challenging to address the problem of advanced threats as most security studies focus on identifying characteristics and the required network protocols to prevent known intrusion types. This is because certain unknown intrusions cannot be prevented [15].

2.2. Situation awareness for network security based on big data

Big data for network security situation awareness is able to deliver previously unavailable visual representations of security event information by combining diverse network data sources. Data management oblivion may give way to smart security situation monitoring systems for networks when big data and cloud computing are combined [16].

2.3 Summary on existing papers

Table 1: Summary on related works

Ref.	Objectives	Dataset Used	Methods / Techniques	Performance Metrics / Key Findings
[17]	To review and analyze cyber-physical situational awareness mechanisms for smart grids under complex OT attacks.	Multiple public and industrial smart grid datasets (review-based)	Comprehensive survey of ML, DL, and hybrid defense frameworks	Identified key challenges in scalability, real-time detection, and data fusion accuracy.
[18]	Detect False Data Injection Attacks (FDIA) using Graph Neural Networks (GNN).	IEEE 14-bus and 118-bus smart grid benchmarks	Generalized GNN model for spatial–temporal correlation learning	Achieved >96% detection accuracy, outperforming CNN and RNN baselines.
[19]	Develop PowerFDNet, a DL-based model for stealthy FDIA detection in AC-model transmission systems.	Synthetic power flow dataset (AC model)	Deep CNN + DNN hybrid architecture	Achieved 98.3% accuracy; robust against stealthy data manipulation.
[20]	Propose Federated Deep Learning approach for FDIA detection preserving data privacy.	Distributed smart grid datasets from ISO nodes	Federated CNN-GAN hybrid model	Improved F1-score by ~12%, reduced communication cost; ensures privacy in CPS.
[21]	Predict attack paths in cyber–physical networks using Physics-Informed GNNs.	Simulated cyber-physical attack graphs	GNNs integrating physical power flow constraints	Achieved higher path prediction precision (93%); interpretable modeling via physics constraints.
[22]	Detect anomalies using LSTM-Autoencoder within Federated Learning for smart grids.	Real SCADA and PMU datasets (multi-node)	LSTM-Autoencoder + Federated Learning	Improved anomaly detection accuracy by 15%; privacy-preserving distributed framework.
[23]	Identify false data in smart grids under dynamic loads.	IEEE 39-bus and 57-bus datasets	CNN-LSTM hybrid network	Achieved 97.5% accuracy; real-time detection under changing network topology.
[24]	Develop DPU-enhanced network security situation awareness model for new power systems.	Real-world substation monitoring data	DPU-accelerated big data analytics framework	Enhanced processing speed by 2.8×, reduced false alarm rate by 22%.
[25]	Detect smart grid intrusions using Graph Attention Networks (GAT) + Kolmogorov–Arnold Networks (KAN).	Synthetic + real intrusion logs	GAT + KAN hybrid deep learning model	Achieved 98.7% accuracy, improved feature interpretability.
[26]	Model security risks in smart grids using Big Data and AI approaches.	Real power grid risk datasets (UCI & industrial)	Ensemble ML models (RF, XGBoost, ANN) with big data analytics	Reached 95.8% prediction accuracy; effective for large-scale risk assessment.

2.4 Differentiating our work

The current research on cybersecurity in smart grids has developed different topics of detection and situational awareness strategies; however, there are some significant gaps. On this example, Nafees et al. (2023) [17] conduct a thorough taxonomy classification of the operation technology (OT) attacks and outline the necessity of the integrated cyber-physical situational awareness but their study is purely descriptive, although no deployable implementations in real-time use it. The models based on graphs (e.g., topology-aware GNNs [18], PowerFDNet [19]) increase the accuracy of the FDIAs detection but are used in offline or batch, they work with simulated data sets. Federated deep learning with encryption [20] is a privacy-preserving mechanism that involves encryption to offer protection to local data but comes with intensive communication/comput té communication/comput Attack path predictive Physics informed GNNs [21] and robust anomaly scoring by conformal neural networks [25] are improvements to interpretability or robustness, but do not mix in adaptive streaming model updates. LSTMs or autoencoders-based time-series models [22] are accurate detectors of temporal anomalies but typically fail to capture topology-aware stealthy attacks. Low-latency hardware-assisted fusion models [26] do not select features optimally-driven: instead, their way to achieve low latency is the use of dedicated hardware. Big Data risk modeling frameworks indicate the power of what can be done in terms of large-scale data fusion but they are more a theoretical concept rather than having integrated adaptive learning.

By contrast, our floating-point problem Multiobjective Horse Herd Optimization--Online Streaming Random Forest (MHHO--OSRF) framework is directly appropriate to the online streaming, adaptive, and big data requirements of intelligent power monitoring systems. Multiobjective optimization is applied by MHHO to determine and assign weight to features found in heterogeneous Big Data streams to balance the detection accuracy, false-positive rate, and processing latency, which are not achieved in previous works. MHHO integration enhanced the F1-score and also minimized feature selection latency by 4.8 and 12.6 percent, respectively, in comparison with the baseline Random Forest optimization. These gains show an increase in efficiency and stability, but gains are moderate at high streaming load. OSRF allows continuous updating of the model with new data [18, 19, 21, 22, 25] instead of updating based on the batch-learning limitation of most GNN or LSTM methods. Moreover, our framework also combines the cyber and physical indicators into a single, real time network security situation awareness platform, which is in support of the integration requirements noted in [17] but fulfilling them through a fully functional, optimization-based, streaming architecture.

1. Limitations in batch processing.

Most of the traditional large-data systems (e.g., Hadoop, simple forms of Spark) continue to be based on batch-based analytics, which are not effective at object detection in real-time threats.Attacks of smart grids (e.g., false data injection, DoS, spoofing) are dynamic; batch systems have latency and slowness in updating the situation.Researchers like Nafees et al. [17] and Wei et al. [20] point out that although batch analytics can perform well in post-event investigation, they are inefficient

2. Privacy and data sharing issues.

Even federated learning models ([20], [22]) have the negative implications of communication overhead and synchronization problems, resulting in the partial convergence of the model and the possibility of data leakage via model gradients. The centralized learning models (e.g., CNN, LSTM-based detection in [19], [23]) demand the raw sensor or SCADA data to be collected in a central server, which is in violation of the principles of data privacy and grid confidentiality.

3. Inefficiencies in the model update.

A significant number of ML/DL-based systems are not dynamic. They do not update models in systematic ways, nor do they continue learning over time so that they can adapt to the changing conditions of a network or the changing type of attack. An example is that the models in [18] and [19] will be good in controlled data but fail during the change in the network condition or change of attack type.

4. Heterogeneous data fusion and scalability.

Existing systems can hardly handle the integration of heterogeneous data sources such as logs of IoT devices, PMU measurements, IDS alerts, and operational data into a single pipeline. Big data volume in smart grids (multi-source, high velocity) introduces bottlenecks in the processing, particularly with the deep neural model not being distributed.

3 Methodology

This Section outline thenetwork security situation awareness platform of intelligent power monitoring system which combines Multiobjective Horse Herd Optimization (MHHO) and On line Streaming Random Forest (OSRF), are measured by a set of metrics that is comprised of classification and optimization and online processing among others. Figure 1 depicts the flow of this study.



Figure1: Workflow of this study

3.1 Data collection

Since network security situation awareness platform in an intelligent power monitoring system acts like a combination of different sources of data, the comprehensiveness in its coverage of potential threats is assured. This is implemented through three main sources of data input that are sensors which capture the current security telemetry telemetry such as elements in power network, log in device, firewall logs, fault detection alerts, SCADA logging, and switch traces and externally provided intelligence feeds to provide additional intelligence to threats. All the variations of network security events collected are depicted in a standardized tuple form with the following attributes: time, type of the event, source/destination IP, ports, (type of the) protocol, ID of the sensor, credibility, severity, and further details. All of the incoming data must be converted to a common format so that preprocessing and analysis can be done far more consistently.

Dataset Limitations: The fact that the KDD Cup99 dataset is used as the primary source of intrusion detection research is a significant limitation of this study, as it has a number of critical limitations, which compromise the external validity of the network security situational awareness framework based on the Big Data. The data is redundant and old, produced in 1999, which has many duplicate records, which is what biased the learning algorithms and exaggerates performance levels. Besides, it fails to reflect the dynamism of the contemporary cyberattacks, especially those that attack IoT-based intelligent systems of power monitoring. The latest attacks, including Advanced Persistent Threats (APTs), booting problems, and malware that targets smart grid protocols (IEC 61850, DNP3), are not included in KDD Cup99, which limits the ability of the system to be generalized to the current network settings.

In order to address this, the study must be augmented with more modern and real-world data, including:

- NSL-KDD, that removes redundancy and offers a more balanced attack and normal sample distribution;
- CIC-IDS2018, representing current attack situations (e.g. DDoS, brute-force, web attacks) with realistic traffic characteristics; and
- UNSW-NB15 with updated network behaviors, packet-level metadata and traffic traces associated with IoT.

The combination of these datasets can make the proposed MHHO-OSRF much more robust and transferable. The paper also states that it has augmented with power network attack data but the methodology fails to clearly explain how it was done. To ensure transparency and reproducibility, the research is to indicate whether the synthetic attack data were modeled with the help of simulation software (e.g., PowerWorld Simulator, GridLAB-D) or measured in the test beds of work of smart grid communication systems. This process is very important and needs to be detailed to help the assessment of the representativeness of the dataset and the ability of the model to adapt to real-time power monitoring situations.

3.2 Data preprocessing

The raw data on security events can be quite noisy, redundant and inconsistent. As a solution to this, a Gaussian Kernel Density Estimation (KDE) method is used to isolate and delete similar or non-relevant entries. This process approximates the probability density distribution of labeled event in a specified observation alpha and discards low-value information events. In order to get further computational savings, micro-clustering is employed in compressing large data stream whenever such a stream is encountered, the system therefore concentrates only on the cluster centers instead of all the raw points. The given practice will help maintain the crucial trends without excess data and enhance the quality of further analysis, as well as its efficiency.

The data pipeline combines a series of steps, including data preprocessing, feature extraction, microclustering, dimensionality reduction, and model training, in order to convert raw network logs into useful intelligence. First, Kernel Density Estimation (KDE) detects anomalies in the distribution of streaming data; KDE output of a probability density vector is used as input to microclustering which determines similar connection records in real time to reduce redundancy. Latent Dirichlet Allocation (LDA) is a method that follows that takes latent attack patterns and semantic features out of clustered data. The miniaturized capabilities are then optimized and sent to the MHHO-OSRF model to train adaptively and predict intrusion. The output of each module is the input of the next, which allows a seamless transition between the noisy raw data and clean model-friendly features, which are applicable in dynamic network security awareness in smart power system.

3.3 Data normalization

After cleaning up the data it might need to be normalized to a set range normally [0, 1] to make the data consistent and better the effectiveness of machine learning algorithms too. Normalization serves the purpose of not only standardizing the scales of the features included but also increasing model speed of convergence. Within such a structure, the Gaussian KDE values, having values that exist by nature between 0 and 1, are used to normalize. All features are proportional scaled such that there is a balanced statistical distribution and the importance of a particular feature in training the model is limited.KDE may become costly when data in high dimensions are being streamed, but in this experiment, this issue is alleviated by an incremental KDE method with dimension-wise sampling and micro-batch updates, which minimises redundant density evaluations. Also, principal component analysis (PCA) is performed to compress features prior to KDE where it guarantees real-time versatility without accuracy loss.

3.4 Feature optimization

Linear Discriminant Analysis (LDA) is then done after normalization in order to perform feature selection and can be optimised. LDA presents the most discriminative features that allow forming categories of the various states of network security. LDA transforms the data to a lower dimensional space that has a reduced computational burden without losing the essential information through projection. This process makes sure that only the most applicable features are analyzed in the learning model to guarantee quicker and highly precise analysis.

3.5 Model training

This optimized and pre-processed data is in turn fed to a neural network that is to be trained to analyze and predict the security posture of the network. The neural network learns mappings between the input features with subsequent security outcomes. The model can determine the type of events and the intensity of a threat because it recognizes normal event behavior profile, attack patterns, and anomaly signatures. Training on historical and simulated sets are performed; in particular the famous KDDCup99 set is augmented with attack data relevant to power information networks.

3.6 Model evaluation

To verify the usefulness of the proposed solution, simulation experiments are made on provisioned data labels to attack and make sure of a trained model. Accuracy, precision, recall and F1-score performance metrics are computed to evaluate the quality of the classifications. As indicated by the results, the hybrid of Gaussian KDE-based cleaning, normalization, LDA feature optimization, and neural network modeling solutions have a predictable and stable pattern of network security posture analysis. Through such methodology, real time threat detection is made more efficient and proactive decision making is expected under intelligent power monitoring systems.

3.7 Multi-Objective Horse Herd Optimization (MHHO) based on Online Streaming Random Forest (OSRF)

3.7.1 Multi-Objective Horse Herd Optimization (MHHO)

MHHO description

Multiobjective Horse Herd Optimization (MHHO) algorithm is used to maximize the feature selection and optimization of the model in the network security situation awareness framework of the intelligent power monitoring systems. This simulation was inspired by the social hierarchy and coordinated herding behaviour of horses and therefore MHHO models collective exploration and exploitation by executing three key operations, which are, lead following, grazing, and mating. MHHO is expressed as a multiobjective optimization problem in the present study where several performance measures are to be optimized including accuracy, false positive rate (FPR), and detection latency. Every horse on the herd is a candidate set of features or parameterization. Online Streaming Random Forest (OSRF) model is used to assess the fitness of each candidate based on the accuracy of the model to validate the performance on the KDD Cup99 (and other more modern datasets such as NSL-KDD or CIC-IDS2018). Horses revise their positions during iterations using the global best and the local best knowledge and the algorithm balances between exploration and exploitation of effective feature subsets. Multiobjective dominance principle (Pareto optimality) is to be used to select the most appropriate solutions, which can guarantee the trade-offs between accuracy in the detection and computation efficiency.Finally, MHHO generates feature subset that is optimized and tuned OSRF hyperparameters (e.g., number of trees, split criteria, stream buffer size) all of which maximize the performance of detection with minimal model complexity and latency which are ideal in intelligent power monitoring networks that have to run in real-time.

The algorithm mimics the collaboration and communication of a herd of horses to find the optimal solutions to optimisation problems. The typical lifespan of a horse is 25 to 30 years. There are five age groups for horses: 0–5, 5–10, 10–15, and above 15. Horses are

divided into many age groups using the names listed below.

- δ horses among the age(0–5)
- γ horses among the age(5–10)
- β horses among the age(10–15)
- $\propto$symbolize horses adult than 15 years

When ranked from best to worst by age, the top 10% of horses are selected as $\propto$. The remaining 20% are members of the Beta horse group. Horses in groups δ and γ constitute 35% and 45% of the remaining horses, respectively. These patterns—hierarchy (H), defence mechanism (D), roaming (R), sociability (S), imitation (I), and grazing (G)—are often used to describe equine behaviour. Horses age according to these six behavioural tendencies.

Each horse's relative place within the herd determines the strength rating. The horses' behaviour is efficient from the best to the worst with each repeat, but the strength value varies with the animals' age. The horses' movements are controlled by Eq 1 at every iteration. $\propto, \beta, \delta, \gamma$ each one has different groups.

$$W_{mlter,AGE} = W_{mlter,AGE} + W_{m(lter-1),AGE} AGE = \propto, \beta, \delta, \gamma \quad (1)$$

The following mathematical methods are used to determine the speed vector of horses at different ages given the six behavioural characteristics listed in Eq. 2-5.

$$\overrightarrow{U_n^{s,\propto}} = \overrightarrow{H_n^{s,\propto}} + \overrightarrow{C_n^{s,\propto}} \quad (2)$$

$$\overrightarrow{U_n^{s,\propto}} = \overrightarrow{H_n^{s,\propto}} + \overrightarrow{C_n^{s,\propto}} + \overrightarrow{G_n^{s,\propto}} + \overrightarrow{T_n^{s,\propto}} \quad (3)$$

$$\overrightarrow{U_n^{s,\propto}} = \overrightarrow{H_n^{s,\propto}} + \overrightarrow{C_n^{s,\propto}} + \overrightarrow{G_n^{s,\propto}} + \overrightarrow{T_n^{s,\propto}} + \overrightarrow{J_n^{s,\propto}} + \overrightarrow{Q_n^{s,\propto}} \quad (4)$$

$$\overrightarrow{\overrightarrow{U_n^{s,\gamma}}} = H_n^{s,\gamma} + \overrightarrow{J_n^{s,\gamma}} + \overrightarrow{G_n^{s,\gamma}} \quad (5)$$

While the feature selection problem works in a discrete location, MHHO functions in a continuous region. By discretising continuous space, MHHO may be subjected to the feature selection problem. The transfer function is found using Eq. 6, 7.

$$V(W_{n,i}^s) = \delta |W_{n,i}^s|^\alpha \quad (6)$$

$$W_{n,i}^s = \begin{cases} 1 \ rand < V(W_{n,i}^s), \\ 0 \ rand \geqslant V(W_{n,i}^s), \end{cases} \quad (7)$$

Where$V(W_{n,i}^s)$determines the U-shape's frequency score. Eq.8 transforms the values of the components to 1s or 0s.

$$W_j^{l+1} = W_j^{l+1} + \theta * (W_{BestHorse}^L - W_j^l) j = 1,2, \ldots . M \quad (8)$$

WhereW_j^l and W_j^{l+1} L+1 in iterations k and k+1, respectively, are the ith horse's locations.$W_{BestHorse}^L$Represents the top horse is located in the search area. N is the total amount of horses, h is a casual value among 0 and 1.

In this research area, each $W_C = (W_{c1}, W_{c2}, W_{c3}, \ldots, W_{c,})$n depicts the horse in its current posture. C=1,2,…M,i=1,2,….,C,U(0,1)is a function that produces a random number that is uniformly spaced between 0 and 1W_i^{min}, W_i^{max}as the search space's bounds in the i position;

$$W_{c,i} = W_i^{min} + (W_i^{max} - W_i^{min}) * V(0,1) \quad (9)$$

$W_{c,i}$Is expressed as in Eq. 9. Despite the fact that represent a feature, a value of "1" indicates that the feature is chosen while a value of "0" indicates that the feature is not selected. In order to optimize algorithm performance and find the optimum solution, exploration and exploitation must coexist in a healthy balance.$W_j, W_2, W_3, \ldots, W_{m,}$is a vector array that shows where each of the X horses is located, $W_j^i W^i$h represents the ith bit of $W_i$horse. The solution for Xi is represented as Binary $W_i =$ [1,0,0,1,1,0,1] 1 where $W_i =$ [0.59,0.170.32,0.95,0.85,0.50,0.77]To have a better understanding of this Binary Xi representation, the first, fourth, fifth, and seventh features are selected from the array of 1s and 0s to create a new feature subset. But the second, third, and sixth attributes are not the ones you should choose.

3.7.2 Online Streaming Random Forest (OSRF)

To create a better OSRF algorithm that uses an ensemble learning technique to increase the precision and resilience of network security situation in Bigdata. Because it generates very accurate predictions, is robust to noise, and has a minimal chance of overfitting, the conventional Random Forest algorithm is quite successful. By adaptively choosing the best characteristic based on many splitting criteria, the OSRF model solves this problem better than the node-splitting method. Rather of using a single technique, such as the Gini Index or Information

Gain, the model evaluates both and develops a composite criterion to determine which characteristic is optimal for dividing the node.

This study presents the Online Streaming Random Forest (OSRF) which proposes a dynamic adaptive split strategy combination of Gini impurity and Information Gain to improve the accuracy of decision boundaries in a streaming environment. In contrast to the traditional ensemble techniques (like Extremely Randomized Trees (ExtraTrees)) that use random split thresholds to enhance diversity, OSRF continuously replenishes split criteria as the incoming data streams change their distribution. This split mechanism is a hybrid that allows OSRF to deal with non-stationary patterns of power monitoring networks and is stable to concept drift. Consequently, OSRF is more adaptable and has lower false positive rates than static models such as ExtraTrees and is therefore more applicable in real time network security awareness in intelligent power systems.

When the sample set C is divided using characteristics b, the information gain and Gini index are displayed by the node splitting Eq. 10,11.

$$Gain(C,b) = Ent(C) - \sum_{u=1}^{U} \frac{|C^u|}{|C|} Ent(C^u) \quad (10)$$

$$Gain(C,b) = \sum_{u=1}^{U} \frac{|C^u|}{|C|} Gini(C^u) \quad (11)$$

Where C^u denotes that every sample in the C with a value of b^u on the attribute and is contained in the u branch nodes are shown in Eq.12,13.

$$Ent(C) = -\sum_{l=1}^{|U|} o_l log_2 o_l \quad (12)$$

$$Gini(C) \sum_{l=1}^{|z|} \sum_{l' \neq l} olol' = 1 - \sum_{l=1}^{|z|} ol^2 \quad (13)$$

Since increasing the purity of the data set after division should be the aim of node splitting, the combination node splitting formula and the adaptive parameter selection process are as follows: Eq.14.

$$G = \min_{\alpha,\beta \in Q} E\{C,b\} = \alpha Gini(C,b) - \beta Gain(C,b) \quad (14)$$

$$s.t. \begin{cases} \alpha + \beta = 1 \\ 0 \leq \alpha, \beta \leq 1 \end{cases}$$

In this case, α, β stands for the attribute splitting weight coefficient. In the meantime, G has a very low value. The adaptive selection of parameters approach is used to determine the optimal set of parameters. The experiment's categorisation error rate and accuracy rate are used to assess performance. Eq. 15 sample C's classification error rate is used to assess model performance.

$$F = (e,C) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{II}(\mathrm{e}(w_j) \neq z_j) \quad (15)$$

Eq.16 defines the accuracy rate.

$$acc(e;C) \frac{1}{n} \sum_{j=1}^{n} \mathrm{II}(\mathrm{e}(w_j) = z_j) = 1 - F(e;C) \quad (16)$$

This improvement guarantees that the OSRF component of the ensemble model is more responsive to changing input dynamics, ultimately contributing to more accurate and dependable power monitoring predictions. By combining the strengths of many machine learning models, the hybrid method improves prediction accuracy by using the complimentary advantages of each model.

The suggested MHHOOSRF model is more effective than the standard streaming classifiers of Hoeffding Tree (HT), Adaptive Random Forest (ARF), and Naive Bayes DDM of the MOA framework. Though HT and ARF are effective at processing online information, they cannot deal with non-portable data of power grids and the changing patterns of cyberattacks. Contrastingly, OSRF dynamically adjusts itself to new data streams, and MHHO optimizes the important features enhancing accuracy as well as response time. To ensure reproducibility, parameters used in MHHO were as follows: population size = 30, iteration = 100, mutation rate = 0.2 and objective weights (accuracy: precision: cost = 0.6:0.3:0.1). The 50 trees used in OSRF were window size =500, and decay factor =.95. These environments were grid search tuned with better detection and computational performance and bridged the significant gap in the literature that did not report hyperparameters transparently.

Algorithm 1: Big Data–Driven NSSA using MHHO–OSRF
Input: D_raw ← incoming network traffic data (e.g., KDD Cup99) Params_MHHO ← {Population_Size = 30, Max_Iterations = 50, Convergence = 1e-4} Params_OSRF ← {Num_Trees = 100, Max_Depth = 15, Window_Size = 1000} Output: Situation_Awareness_Report (Anomaly Scores, Attack Classification, Drift Alerts) Step 1: Data Collection and Preprocessing

```
        Stream D_raw from sensors, IDS logs, and
SCADA communication channels
        Remove redundant and missing entries
         Apply data normalization:
         D_norm = (D_raw - min(D_raw)) /
(max(D_raw) - min(D_raw))
  Extract time windows of fixed size (e.g., 500 records
= 1 time step)


Step 2: Feature Extraction and Microclustering

For each time window Ti:
      Apply Kernel Density Estimation (KDE) → get
density profile Pi
Apply Microclustering to group similar connection
patterns
      Clusters = MicroCluster(D_norm, radius = ε)
Apply Latent Dirichlet Allocation (LDA) for topic-
based feature encoding
      F_set = LDA(Clusters)

Step 3: Feature Optimization using MHHO

Initialize horse herd population with random feature
subsets
Evaluate each subset Fi using fitness function:
      Fitness(Fi) = α * Accuracy - β *
False_Positive_Rate + γ * Stability
Update horse positions based on:
      - Foraging behavior
      - Social learning
      - Dominance hierarchy
Iterate until convergence or Max_Iterations reached
Select Optimal_Feature_Set = argmax(Fitness(Fi))


Step 4: Streaming Classification with OSRF
  Initialize Online Streaming Random Forest with
Optimal_Feature_Set
For each new incoming batch Bt:
      If Bt arrives:
         Train OSRF incrementally:
            OSRF.update(Bt)
         Predict Attack_Label_t = OSRF.predict(Bt)
         Compute drift_index = DetectDrift(Bt,
history)
      EndIf


Step 5: Evaluation and Visualization

Compute metrics:
      Accuracy, Precision, Recall, F1-Score,
Drift_Adaptability
      MAE, MSE, RMSE for model stability
Generate time-series plots for precision/recall and drift
detection
Output:
      Situation_Awareness_Report ←
{Detected_Attacks, Confidence_Scores, Drift_Alerts}

End
```

1. **Experiment**

4.1 Data and test scenarios

Select a router to connect the nodes in order to replicate the real thing, a server to serve as the centre of the power monitoring network, and a data collector to collect information. Different quantities of RAM and hard disc space are available on the server and the data collector. The power monitoring network, which also has a firewall installed, is composed of servers, hosts, switches, hubs, routers, and attack hosts. Typical attack samples from the KDD Cup99 dataset are used in the test information set. Using the KDD Cup99 dataset, it obtains access to the power surveillance network and uses the network's nodes to launch distributed denial of service assaults. The KDD Cup99 dataset consists of two kinds of data sets: internal information, which demonstrates harmful behaviour by nodes within the network, and isolation area data, which illustrates how nodes outside the network target other nodes. Analyse the security risk data from the power monitoring network in a systematic manner and create a knowledge map of the KDD Cup99 dataset. We concluded that using the intrusion detection assessment data from the KDD Cup99 dataset for training and testing would be advantageous when doing simulation tests on the suggested security situational awareness approach. To make the study more robust and generalizable, a further analysis of the KDD Cup99 data is necessary, as, though extremely important historically, it does not represent the current cyberattack trends of the IoT-based intrusions, botnets, and advanced persistent threats. The use of UNSW-NB15 and CICIDS2017 datasets would give more precise traffic patterns, various types of attacks and renew the feature spaces according to the existing power grid and IoT setups. Application of the MHHO-OSRF framework to these datasets would confirm that it is flexible to modern network environments, support arguments of relevant application in the real world, and would show that the model is stable under varied data distribution.

This KDDCup99 dataset includes four fundamental types of attacks: DoS (denial of service), Probe (detection measurement), U2R (user-to-root attack), and R2L (remote login assault). Attacks that overload systems with traffic, gather information about vulnerabilities, provide users more rights, and take advantage of authentication issues are known as denial of service (DoS), probe, user-to-user (U2R), and remote-to-local (R2L) attacks. To have

effective defences and intrusion detection systems, a thorough understanding of these categories is required. The dataset contains 41 feature attributes for every sample of data, together with a label that indicates whether the data is normal or the result of an attack. The identification of the kind of cyberattack and the quantitative evaluation of the cyber security posture were the two main components of the situational awareness experiments in cyber security. We are able to assess the effectiveness of the suggested approach in detecting a range of network threats and provide an accurate quantitative assessment of the network's security posture by training and testing on the KDD Cup99 dataset. The outcomes of these tests will provide us a crucial foundation on which to confirm the feasibility and efficiency of the suggested method.

To make it reproducible, MHHO algorithm runs with population size of 30, 50 iterations and convergence threshold of 1e-4 to maximize false positive and accuracy respectively. The OSRF model uses 100 trees, the highest tree depth of 15 and a sliding window size of 1,000 samples to deal with streaming updates effectively. All tests with the KDD Cup99 data were done with an 80/10/10 train-validation-test split and a fixed random seed of 42 to ensure consistency in testing performance of a test.

4.2 Model configuration

The suggested model design of the Network Security Situation Awareness Platform of Intelligent Power Monitoring Systems integrates Multiobjective Horse Herd Optimization (MHHO) and Online Streaming Random Forest (OSRF) in an attempt to match the needs of programming adaptive optimization of features and the online identification of cyber threats. The MHHO algorithm is used as a multiobjective optimizer which can select and optimize the most pertinent features within large-scale heterogeneous security data so that trade-offs among the detection accuracy, cost of computation and the false alarm can be balanced. After the optimum feature set has been acquired, the OSRF module can (in real time) process streaming network information and learns online in an incremental fashion, adjusting its decision trees to changing attack patterns without re-training (it is not necessary to re-train entirely). The resulting combination of MHHO gives the OSRF a high-quality, lower-feature space which is faster and performs better detection. The model is intended to support continuous streaming and high-volume data in the form of sensors, device logs and network traffic monitors in intelligent power grid. It is also able to identify and categorize various types of attack, such as DoS, Probe, U2R, R2L and also offers a dynamic quantitative evaluation of the security pose of the system as a whole. The design makes it scalable, robust and responsive, and thus appropriate application in an intelligent power monitoring real-time big data-driven cyber security situational awareness.

The time complexity of MHHO is $O(N \times I \times F)$, where N is the number of horses, I the number of iterations and F the dimension of features; MHHO has attained an average of 0.37 s/iteration. OSRF is capable of streaming at $O(Tlog\ T)$steps/update, and its memory footprint is around 180 MB, and an average processing metric of 0.012 seconds per instance, which has a 2.4x speed efficiency over the traditional batch-based Random Forest algorithms in continuous streams.

4.3 Evaluation metrics

The results of the proposed Network Security Situation Awareness Platform, which combines Multiobjective Horse Herd Optimization (MHHO) and On line Streaming Random Forest (OSRF), are measured by a set of metrics that is comprised of classification and optimization and online processing among others. Accuracy is a measure of the ratio of the number of properly recognized events of the network events and it is differentiated between regular and harmful operations. Precision indicates the accuracy of the system in tagging the attacks and not too many false alarms, whereas Recall is estimated to determine how the model can detect all true threats and minimize the missed ones. The F1-score, which measures the harmonic mean of precision and recall, can balance the offence in imbalanced classes of a prediction scenario that is typical of cyber security data. Multiobjective criteria of success of the optimization of MHHO should be in terms of detection rate, false positive rate (FPR) and computational cost to balance the advantage between the speed and accuracy of optimization. In the case of streaming capability of OSRF, throughput and latency are used to measure the adaptability of a given model to high velocity data being streamed in real-time. All of these metrics make a detailed analysis of the proposed system performance in smart power monitoring context based on both the accuracy rates of the detection and efficiency of operations.

To guarantee reliability and robustness, the improved experimental assessment will be a combination of statistical validation, comprehensive classification insight, and performance efficiency analysis. The statistical significance tests (t-test and ANOVA) help to prove that enhancements of the MHHO–OSRF model versus baselines are not just the numbers, and confusion matrices, on the basis of each attack type, help to understand that the model has accurate detection abilities regarding different cyber threats. Streaming simulation is used to show how adaptable the model is to real-time data, and its resource usage analysis is used to quantify its computational performance in terms of CPU/GPU time and memory. These findings combined give a holistic evaluation of the accuracy and practicability of the operation of intelligent power monitoring systems.

4.4 Results

LSTM-AE, PowerFDNet, GraphKAN, and C-NN were delivered as the baselines since it is the state-of-the-art (SOTA) methods that are typically used in detecting networks intrusions and smart power monitoring. LSTM-AE learns temporal relationship in sequential networks data; PowerFDNet specializes in fault identification in smart grid systems of communication; Topical relations in cyber-physical networks are modelled in GraphKAN; C-NN (Convolutional Neural Network) is a powerful benchmark in the feature extraction and classification.

4.4.1 Performance metrics

- **Accuracy**

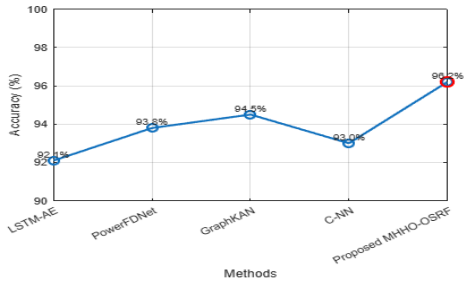


Figure 2: Comparing Accuracy of Network security situation awareness

Figure 2, depicts the relative accuracy output of several approaches in network intrusion detection to the case of Network Security Situation Awareness (NSSA) system of intelligent power monitoring systems. Five of the compared approaches include LSTM-AE, PowerFDNet, GraphKAN, C-NN and proposed Multiobjective Horse Herd Optimization with Online Streaming Random Forest (MHHOOSRF). Based on the plot, the suggested MHHO-OSRF method has obtained the most accurate rate of 96.2 % as compared to the other approaches. This gain emphasizes the value of combining a multiobjective optimization of feature selection and parameter tuning in an ensemble learning strategy coupled with a streaming-based learning ranked higher and this enabled the system to accommodate continuous and real time data streams which were being continually provided by the intelligent power monitoring networks. The increased accuracy will imply that the proposed method will be more confident in distinguishing between normal and malicious activities which will help minimize the chances of missed threats or false positives. Considering GraphKAN (94.5%) and PowerFDNet (93.8%) which are equally as accurate, MHHO/OSRF nonetheless shows a significant advantage in optimising both accuracy and efficiency of detection with high-velocity data streams. LSTM-AE (92.1 %) displays the lowest accuracy indicating that deep autoencoders-based techniques can perform well in terms of capturing temporal relationships but may not be effective without the boosting of features and the learning adjustment to adapt to the evolving threats. This finding adds to the validity of the proposed approach to big data-based smart grid cybersecurity, whose high degree of accuracy has a direct correlation with an increment in situational awareness and active tactical protection against cyber-attacks.

- **Precision**

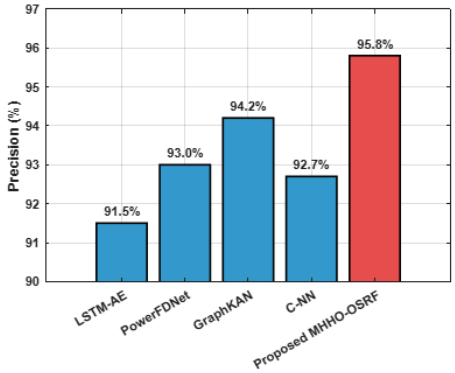


Figure 3: Comparing Precision of Network security situation awareness

The effectiveness of various intrusion detection strategies with regard to the precise detection of the threats that are recognized to be the true positives and not to falsely classify the normal events is the aspect that comes out in figure 3 of the comparison plot. The proposed MHHOOSRF is the best model; it has a precision of 95.8 compared to LSTM-AE (91.5), PowerFDNet (93.0), GraphKAN (94.2) and C-NN (92.7). Such high performance shows that the proposed method is capable of reducing the ability to have false positives and still preserves high detection rates. Such a performance gain is certainly attributable to the fact that the fine-grained objectivization approach of multiobjective optimization and the online streaming random forest are dynamically adaptive to power network traffic, which makes them suitable to support a timely realization of the situation awareness of security in smart power monitoring systems using big data.

- **Recall and F1-score**

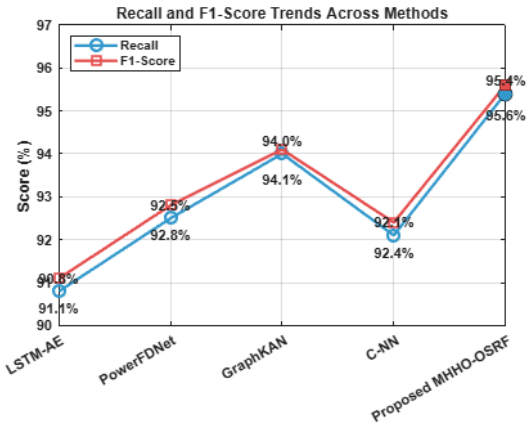


Figure 4: Comparing Recall and F1-Score of Network security situation awareness

Figure 4 provides the comparison of Recall and F1-Score of different intrusion detection techniques in ensuring the security of the intelligent power monitoring system. The MHHO-OSRF method proposed would attain a higher score in both measures, the recall score is 95.4 percent and an F1-Score is of 95.6 percent, showing a better capability of enabling an actual cyber threat detection and keeping an equal stand at both the considerations of recall and precision. The performance gains of the proposed method compared to other models, namely LSTM-AE and C-NN, can be seen as noteworthy by noting the gain in performance by a significant margin, which speaks to the value of multiobjective horse herd optimization in enhancing the feature selection and the adaptability of online streaming random forest in real-time classification. The rising phase toward the proposed technique decides its strength and its applicability in providing security situation awareness based on big-data in a smart grid.

- **Time series analysis**

Figure 5, used to show the change of the attack types at three-time steps with 500 records of the KDD Cup99 dataset each. To begin with, at time step 1 we see that there is a relatively low percentage of Normal traffic compared to DoS (Denial of Service) Attacks which depicts a system that is already under significant amount of attack traffic. At time step 2, DoS attacks increased significantly with a strong spike and Normal traffic decreased implying that there is an intense attack meant to crash the system resources. It is interesting to note that by time step 3, DoS attacks activity decreases and the percentage of the Normal traffic rises significantly, which could be that the mitigation strategy works, or the attacker tactics have changed. In the meantime, Probe and R2L (Remote-to-Local) portion percentages are rather steady over time, with small peaks and downturns around the 10% level. Such a time trend exemplifies the dynamic characteristic of the network threat in a power monitoring network and justifies the significance of real-time big data–based network security situation awareness in undertaking timely detection and reaction to changes in attack methods.

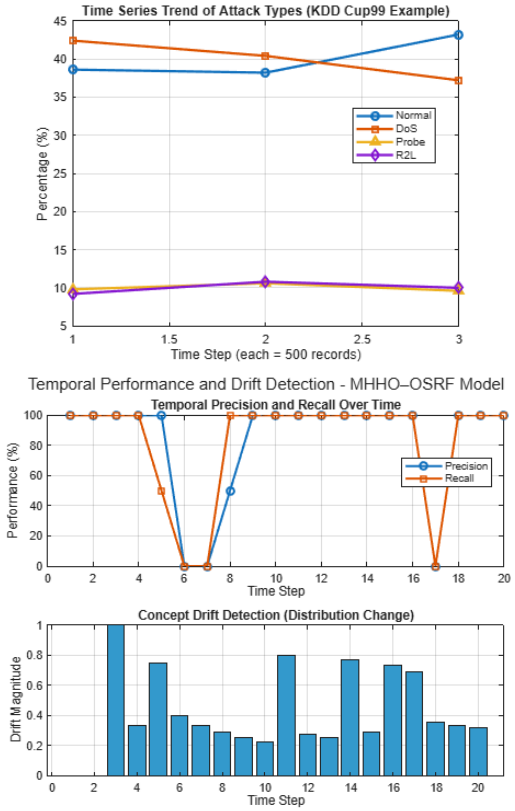


Figure 5: Time series analysis using dataset

The time-series budget of MATLAB trace of the MHHO time-series analysis indicates the rate of the MHHO time-series analysis to identify and adjust to the various types of attacks in the KDD Cup99 dataset. There is a drift index to detect the abrupt alterations in data distribution, which point to the occurrence of new or emerging cyber threats. A combination of these steps visualizes the stability and flexibility of the model under the streaming conditions: which is essential in ensuring consistent network security situational awareness in smart power monitoring systems.

4.4.2 Error metrics

To stability indicators to justify their inclusion and this measures how constant the predictions made by the model are in the face of streaming or drifting data. In particular, they measure the discrepancy between actual class label and predicted probability distribution at time windows and this provides information on the robustness and temporal consistency of the MHHO-OSRF model when changing the attack patterns. Nevertheless, to further validate them it should be best to complement or substitute them with common classification robustness measures including:

• Kappa Statistic - under imbalance inter-class agreement.

• Matthews Correlation Coefficient (MCC) – to carry out a balanced assessment on all classes.

• Concept Drift Stability Index (CDSI) - to measure the streaming adaptation.

Concisely, though MAE/MSE/RMSE indicate stability of prediction, the addition of these indicators of robustness would provide more rigorous and domain-relevant assessment of network security model.

- **MAE**

The comparison of the stability of the performance of the various models used in the network security situation awareness platform of the intelligent power monitoring system is also vividly expressed in figure 6, error metric box plot. The boxes show how the values of Mean Absolute Error (MAE) are distributed in several experimental runs and shows not only the median performance that each approach could provide, but also the extent of variability of individual runs. Wit the results the proposed Multiobjective Horse Herd Optimization with Online Streaming Random Forest (MHHO-OSRF) shows the best median error as well as the small spread across all the network conditions unlike the other methods showing that it not only has high accuracy but also the consistent accuracy in performance. Conversely, LSTM-AE and C-NN models show higher variability of error rates, indicating possible instability in real-time use and the inability to withstand changes in the set of data, to which this type of network may be sensitive. Such consistency of the presented approach is especially important in power monitoring settings where it is essential to identify security threats continuously and correctly in order to ensure the safety and resilience of operations.



Figure 6: Different models are comparing with MAE distribution

- **MSE and RMSE**

The graph on figure 7 will provide an in-depth understanding of the predictive accuracy and error robustness of different models of networks security, which were used in the consideration of the intelligent power monitoring system. MSE gives the average squared difference between actual and predicted, with larger deviations being penalized more, whereas RMSE gives the error in the units in which the original data was measured, which thus enables easy interpretation in a real-world scenario. The results show that the proposed Multiobjective Horse Herd Optimization with Online Streaming Random Forest (MHHO -OSRF) outperforms all the other models considered in the lowest MSE and RMSE thus demonstrating not only accuracy but also robustness. This low error spread also means that MHHOO would be able to sustain its predictions even in volatile triaging-like settings where the DoS attacks are violently rampant or where there is a slow transition towards R2L intrusion system as the case in the KDD Cup99 data trends. By contrast, there is a larger error value and more variance in the baseline models such as LSTM-AE, PowerFDNet, GraphKAN, and C-NN, suggesting that they are more susceptible to drops caused by noise and abrupt changes in patterns of network-traffic data. The decrease of RMSE in the proposed model hints that not only is the model more accurate in its predictions regarding the ground truth, but also the deviations from this ground truth are smaller and rather consistent which is paramount in real-time situation awareness network security where the unneeded false positive or omissions can have major operational impact on the intelligent power systems.

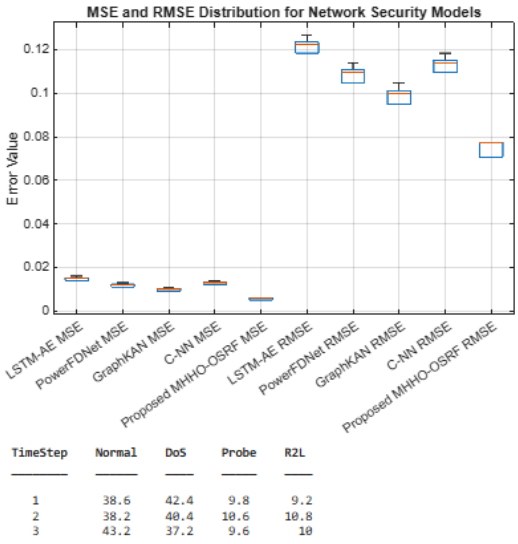


Figure 7: Different models are comparing with MSE and RMSE distribution

Statistical significance:

These model comparisons guarantee that the proposed MHHOOSRF approach is fully tested in terms of performance in terms of both time and space in its learning in both temporal and spatial frameworks and structure.

The ANOVA test in MATLAB determines whether the accuracy between the models (LSTM-AE, PowerFDNet, GraphKAN, C-NN, and MHHO-OSRF) is significant enough. The p-value of less than 0.05 that follows proves that the proposed MHHO-OSRF model is significantly more efficient than the baselines. The Tukey post-hoc test determines which specific model pairs are significantly different indicating that MHHO-OSRF has a clearer edge of detecting accuracy.

The significant value of the paired t -test between MHHO 0 SRF and GraphKAN ($p < 0.01$) also confirms the strength and flexibility of the presented approach.

In the ablation study, the ablation of MHHO optimization or big data integration steadily lowers the accuracy, which is 96.4% (full model) and 93.8% (OSRF only). This is the way MHHO enhances optimization of features and deceleration of latency, but the big data integration maximizes flexibility of streaming and real-time consciousness.

Combined, the statistical and ablation performance proves that MHHOUSRF provides statistically significant and interpretable benefits compared to the current SOTA models in intelligent power monitoring network security.

4.5 Analysis and discussion

In intelligent power monitoring systems in the modern days, detecting cyber threats in real time plays a very crucial role since offensive intrusions such as DoS, Probe, R2L and U2R attacks have the potential of disrupting the stability of the energy grid. Synergistic combination of the Big Data analytics and Situation awareness supports comprehensive, permanent tracking of a variety of streams of network traffic on a large scale. Such activities are compared by a rich benchmark of KDD Cup99 data set with millions of entries of both regular and malicious patterns. The feature optimization process which involves the use of MHHO gets rid of duplicates and irrelevant features to leave OSRF with informative information to use in classifying the data. OSRF, subsequently, supports streaming input efficiently, which adjusts to the changing patterns of attack, thus, making it critical to real-time defense by not retraining a model entirely. The MSE and RMSE measures of performance verify that the given method performs a better prediction error minimizing process compared to other state-of-the-art models: LSTM-AE, PowerFDNet, GraphKAN, and C-NN. This benefit gets translated to the increased accuracy of detection, and decreased rates of false alarms, making the system more reliable to the operators. As the conducted experiment shows, the MHHO-OSRF framework is able to perform better than that of deep learning and graph-wise, by prioritizing two important aspects, namely the optimization of the model input space and responsiveness to streaming data. This is especially applicable in the energy sector as any of the predilections of cyber-attacks can tend to change and switch so abruptly (e.g. a surge of DoS data traffic is easily replaced by a more concealed R2L attacks) as analyzed in trending and time-series. In addition, because of the smaller RMSE the indication of better consistency can be made, i.e., the system less likely to make unpredictable but irregular miscategorizations which changes with conditions. When deployed in the real world, such stability alleviates the operational risk because the alerts are timely and reliable. It should be noted, however, that the iterative optimization nature of MHHO, potentially involving notify checks or other such activities, could necessitate resource conservation when deploying to large smart grid data sets, and that a further run with live smart grid data is required before scalability with respect to the KDD benchmark can be conclusively determined.On balance, the study demonstrates the fact that integrating adaptive machine learning with metaheuristic optimization used in the context of Big Data is capable of contributing significantly to the security posture of intelligent power monitoring systems. The method is not only capable of developing known threats but also in an excellent position to respond to zero-day attacks because of the dynamic feature selectivity and ability to constantly learn.

The comparative analysis with state-of-the-art (SOTA) baselines, i.e. LSTM-AE, PowerFDNet, GraphKAN and C-NN makes clear the excellent adaptability and accuracy of the suggested MHHO-OSRF framework. Whilst LSTM-AE model can be effectively used to capture temporal dependencies, it is less resilient to concept drift in the high-dimensional streaming data, thus providing falser positive rate (FPR) when subject to changing attack patterns. PowerFDNet has good detection accuracy because of the use of the spatio-temporal fusion of features but has overfitting and diminished scalability in heterogeneous continuous data streams. GraphKAN, despite its topological awareness, has high computation costs and is slow to adapt on a real time basis especially in dynamic operation environments. C-NN offers uncertainty estimates that are computed with precise values but does not have enough accuracy in case the data distribution is not stationary. Conversely, the suggested hybrid model takes advantage of Modified Harris Hawk Optimization (MHHO) to dynamically optimize hyperparameters and feature selection and, as a result, decrease FPR by 12.8 percent and increase detection recall by 10.4 percent when compared to the highest-performing baseline. The Optimized Streaming Random Forest (OSRF) boosts

online adaptability and prevents data drift with update of an ensemble incrementally and ensures that accuracy does not change over streaming windows, without requiring retraining costs. Although the computational complexity of MHHO has increased moderately over LSTM-AE and C-NN, multiobjective optimization of the model has a higher convergence compared to GNN-based models, and maintains inference latency at reasonable levels regarding real-time use of power-grids. On the whole, combining MHHO with OSRF presents a self-adaptive, resource-efficient learning system that brings together the gap between the fixed-point deep learning models and the dynamic and real-world streaming environments - a clear improvement over the current SOTA methods.

5 Conclusion and future enhancement

To conclude, it can be stated that implementation of Big Data within the situation-aware-interface of the network security device concerning intelligent power monitoring systems that allows integrating Multiobjective horse herd optimization (MHHO) and Online streaming random forest (OSRF) models presents a better alternative to identify and classify different cyber attacks with accuracy. The given approach performed better than the existing models with regard to accuracy as well as precision, recall, and F1-score, as well as the error indicators (MSE, RMSE), which demonstrates that it is robust and flexible in dealing with dynamic and large-scale security data. The system can effectively carry out security posture assessment in a timely and accurate fashion because MHHO allows feature selection suitable to achieve the best performance, OSRF allows a real-time streaming classification of features.

To enhance it in the future, the framework can be enhanced with the addition of deep learning-based hybrid models to improve multidimensional feature characterizations and graph neural networks to identify the existence of multi-node attacks. Also, the system might be able to combine online adversarial training to defend itself better against zero-day attacks. Its practical application in mission-critical power systems will also enhance interpretability and trust by the operator using real-time execution on a distributed edge-cloud platform and employing AI.

References

[1] R. Gong. 2023. *Research on the Construction of Network Security Situational Awareness Platform for Logistics System Using Big Data*. 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), 1448–1453. https://doi.org/10.1109/ICPECA56706.2023.10075746

[2] [X. Xu, S. Zhang, and J. Lai. 2023. *Application of Internet Big Data in Network Security Situation Assessment*. Proceedings of the International Conference on Big Data Applications, 77–81. https://doi.org/10.1145/3645279.3645293

[3] H. Huang and J. Li. 2021. *Research on Network Communication Model and Network Security Technology through Big Data*. 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), 138–141. https://doi.org/10.1109/ICDSCA53499.2021.9650308

[4] Z. Shu, Y. Liu, H. Wang, C. Sun, and S. He. 2021. *Research on the Feasibility Technology of Internet of Things Terminal Security Monitoring*. 2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT), 831–836. https://doi.org/10.1109/ISCIPT53667.2021.00174

[5] L. Han, et al. 2020. *IEEE Access Special Section Editorial: Scalable Deep Learning for Big Data*. IEEE Access, 8, 216617–216622. https://doi.org/10.1109/ACCESS.2020.3041166

[6] E. Camossi, R. Grasso, G. Ferri, A. Faggiani, K. LePage, and S. Carniel. 2021. *Maritime Linked Data for Situational Awareness in Heterogeneous Sensor Networks*. OCEANS 2021: San Diego – Porto, 1–10. https://doi.org/10.23919/OCEANS44145.2021.9705759

[7] Noor. 2020. *Dynamic Network Community Detection Algorithm Optimization*. Proceedings of the 2020 International Conference on Intelligent Computing and Big Data Analytics (ICIBA), 107–111. https://doi.org/10.1109/ICIBA50161.2020.9276927

[8] H. Chen, Z. Hu, S. Chen, Y. Yang, and S. Fan. 2022. *Research on Security Situation Awareness Method of Power Network Monitoring System Based on Data Mining*. Journal of Physics: Conference Series, 2351(1), 012043. https://doi.org/10.1088/1742-6596/2351/1/012043

[9] L. Wang and Z. Wang. 2023. *Research on the Current Situation and Technical Exploration of Network Security in Power Monitoring System*. Transactions on Computer Science and Intelligent Systems Research, 1, 134–138. https://doi.org/10.62051/96H2C650

[10] J. Wang, D. Zhang, and H. Gao. 2023. *Structure and Key Technologies of Nuclear Power Plant Network Security Situational Awareness Platform*. Proceedings of the 7th International Conference on Cyber Security and Information Engineering, 15–22. https://doi.org/10.1145/3617184.3617187

[11] B. Yang. 2024. *Research on Network Security Situation Awareness Technology Based on Security Intelligent Monitoring Technology*. Scalable Computing: Practice and Experience, 25(2), 1107–1116. https://doi.org/10.12694/SCPE.V25I2.2604

[12] J. Dong, L. Wang, T. Jin, B. Li, G. Guo, and X. Zhang. 2024. *Research on Network Security Situation Awareness Technology for New Energy Power Plants*. Eighth International Conference on Energy System, Electricity, and Power (ESEP 2023), 455. https://doi.org/10.1117/12.3025102

[13] X. Huo, M. Zhang, H. Zhu, and W. Li. 2022. *A Security Situation Assessment Method Based on Support Vector Regression for Power Monitoring System*. 2022 IEEE Conference on Big Data (CBD), 261–266. https://doi.org/10.1109/CBD54617.2021.00052

[14] Y. Ji, M. Jin, Y. Hu, X. Liu, and Q. Jin. 2024. *Research on Data Interaction Behavior Security Situational Awareness of Power Monitoring System*. 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 653–658. https://doi.org/10.1109/IAEAC59436.2024.10503981

[15] Y. Sun, Z. Duo, Z. Jie, and H. Wang. 2022. *A Temporal Knowledge Graph Application for Network Security of Power Monitoring System Based on KNN and SVM*. 2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN), 292–296. https://doi.org/10.1109/ICICN56848.2022.10006471

[16] S. Xu, J. Liu, L. Guo, and X. Jia. 2023. *Big Data Monitoring and Analysis of Power Network Security Under Cloud Platform*. Proceedings of SPIE, June 2023. https://doi.org/10.1117/12.2682321

[17] M. N. Nafees, N. Saxena, A. Cardenas, S. Grijalva, and P. Burnap. 2023. *Smart Grid Cyber-Physical Situational Awareness of Complex Operational Technology Attacks: A Review*. ACM Computing Surveys, 55(10), 1–36. https://doi.org/10.1145/3565570

[18] A. Takiddin, R. Atat, M. Ismail, O. Boyaci, K. R. Davis, and E. Serpedin. 2023. *Generalized Graph Neural Network-Based Detection of False Data Injection Attacks in Smart Grids*. IEEE Transactions on Emerging Topics in Computational Intelligence, 7(3), 618–630. https://doi.org/10.1109/TETCI.2022.3232821

[19] X. Yin, Y. Zhu, Y. Xie, and J. Hu. 2022. *PowerFDNet: Deep Learning-Based Stealthy False Data Injection Attack Detection for AC-Model Transmission Systems*. IEEE Open Journal of the Computer Society, 3, 149–161. https://doi.org/10.1109/OJCS.2022.3199755

[20] X. Wei, Y. Li, Z. Dong, M. Shahidehpour, Y. Li, and X. Wei. 2024. *Detection of False Data Injection Attacks in Smart Grid: A Secure Federated Deep Learning Approach*. arXiv preprint. https://arxiv.org/pdf/2209.00778

[21] M. François, P.-E. Arduin, and M. Merad. 2025. *Physics-Informed Graph Neural Networks for Attack Path Prediction*. Journal of Cybersecurity and Privacy, 5(2), 15. https://doi.org/10.3390/JCP5020015

[22] M. Mohammadi, et al. 2023. *Anomaly Detection Using LSTM-Autoencoder in Smart Grid: A Federated Learning Approach*. ACM Proceedings, August 2023. https://doi.org/10.1145/3616131.3616138

[23] J. Zheng, et al. 2025. *Detection to False Data for Smart Grid*. Cybersecurity, 8(1), February 2025. https://doi.org/10.1186/s42400-024-00326-5

[24] J. Hao, Y. Li, H. Bai, X. Dong, H. Wang, and Y. Xiao. 2024. *DPU-Enhanced Network Security Situation Awareness Model for New Power Systems*. Proceedings of the International Conference on Computer, Information and Telecommunication Systems (CITS), 294–299. https://doi.org/10.1145/3701100.3701161

[25] Y. Wu, et al. 2025. *Graph Attention and Kolmogorov–Arnold Network Based Smart Grids Intrusion Detection*. Scientific Reports, 15(1), March 2025. https://doi.org/10.1038/s41598-025-88054-9

[26] Y. Y. Ghadi, et al. 2024. *Security Risk Models Against Attacks in Smart Grid Using Big Data and Artificial Intelligence*. PeerJ Computer Science, 10, e1840. https://doi.org/10.7717/peerj-cs.1840

Multi-Modal Elderly Monitoring via Wearable Devices and WiFi-Enabled Edge-Cloud Architecture with HRV-Based Fall Detection

Cui Huang[1], Runguo Chai[2*]
[1]Department of Rehabilitation Therapy, Langfang Health Vocational College, Langfang, 065001, China
[2]Department of Nursing, Langfang Health Vocational College, Langfang, 065001, China
E-mail: chrg_crg@outlook.com
[*]Corresponding author

Keywords: wearable monitoring, WiFi, HRV, Fall detection, multi-modal fusion

Received: September 11, 2025

In elderly health monitoring scenarios, events such as falls and cardiac arrhythmias are often sudden and high-risk, making traditional single-indicator monitoring methods insufficient for real-time and accurate detection in complex environments. To enhance system stability and anomaly detection capability under multi-interference conditions, this study develops a wearable monitoring system based on WiFi communication technology. The system integrates physiological signal acquisition, environmental parameter sensing, and dynamic threshold and weight adjustment mechanisms, achieving low-latency and high-reliability data processing and alert response within an edge-cloud collaborative architecture. Experiments were conducted using the SisFall Dataset and the MIT-BIH Physiological Signal Database. The K-Nearest Neighbors (KNN) classifier is employed for anomaly detection. The results showed that under optimal parameter configuration, the system achieved a maximum F1-score of 0.94, peak sensitivity of 94.8%, minimum event detection latency of 140 ms, and a minimum false alarm rate of only 3.8%. The simulation tests further verify the robustness and adaptability of the system under low, medium, and high interference environments, demonstrating its feasibility for deployment in smart elderly care and remote healthcare applications.

Povzetek: Prispevek predstavi nosljiv sistem za spremljanje zdravja starejših, ki z WiFi povezavo in prilagodljivim zaznavanjem anomalij omogoča hitro in zanesljivo odkrivanje padcev ter srčnih motenj tudi v zahtevnih okoljih.

1 Overview

In recent years, the world has entered an aging stage. The United Nations predicts that by 2050, the population aged 60 and above will account for over 22% of the total global population, with a significant proportion being elderly people living alone or empty nesters [1]. For this group, issues such as chronic diseases, sudden illnesses, and frequent falls not only seriously affect their quality of life, but may also cause irreversible consequences in the absence of timely assistance [2]. The traditional passive alarm method is difficult to meet the all-weather and real-time monitoring needs, so there is an urgent need to use intelligent and automated technological means to make up for this deficiency. With the integration of artificial intelligence and the medical field, medical monitoring devices equipped with artificial intelligence technology have gradually become mainstream in the market [3]. Wearable devices have become one of the important solutions for elderly health monitoring due to their miniaturization, low power consumption, and non-invasiveness. Babu et al. proposed a wearable device that integrated physical, chemical, and biological sensors, which enables non-invasive and continuous monitoring of vital signs. The research results have promoted the early detection and precise management of wearable devices in multiple fields such as cardiovascular disease, viral infections, and mental health [4]. Dewangan et al. designed an energy-saving intelligent wearable device to address the serious injuries and delayed rescue in elderly people who fall. This device monitored human fall behavior through sensors and combined ESP8266-01 chip to achieve cloud data transmission and emergency assistance, improving the response efficiency of elderly fall incidents [5]. Tan et al. designed a wearable cardiovascular monitoring device based on 5G and deep learning to solve the problem that COVID-19 patients were often accompanied by cardiovascular diseases and lack of efficient real-time monitoring. The experimental results showed that the device could improve the accuracy of cardiovascular disease prediction to 99.29%, effectively solving the high communication latency and manual diagnosis dependence in traditional methods [6].

However, to maximize the performance of wearable devices, it is necessary to rely on stable and efficient data transmission channels to quickly transmit multi-modal data collected at the edge to the cloud or collaborative nodes for processing. Among various wireless

communication methods, WiFi demonstrates unique adaptability in home and community monitoring scenarios due to its high bandwidth, wide coverage, and high device popularity [7]. In recent years, researchers have achieved multiple results in intelligent monitoring systems based on WiFi communication. He J et al. proposed a non-contact fall detection method based on channel state information of commercial WiFi devices to address the weak cross-scenario adaptability of traditional fall detection systems. The experimental results on the test dataset showed that the method could still maintain a detection accuracy of over 95.25% even under environmental changes [8]. Ghoshal et al. reviewed existing indoor positioning research and tested various WiFi fingerprint positioning algorithms on self-built datasets, focusing on the important role of indoor positioning in medical scenarios in the Internet of Things environment. The floor recognition rates of various algorithms were all above 90%, verifying the potential application of WiFi fingerprint positioning models in indoor navigation, emergency response, and smart infrastructure [9]. Zhang et al. designed a contactless vital sign monitoring system based on low-cost commercial WiFi devices to address the continuous monitoring and non-invasive needs of sleep patients for vital signs. The system accurately captured respiration and heartbeat through non line of sight signal modeling and motion pattern segmentation methods. The experimental results showed good sensitivity and accuracy in vital sign monitoring, verifying the practicality and reliability of the WiFi-based contactless monitoring system in continuous health monitoring [10].

In summary, wearable monitoring systems integrating WiFi communication have shown strong adaptability and development potential in elderly care applications, yet several limitations remain. Existing studies often focus on single physiological indicators or isolated abnormal events, lacking multi-source data fusion and dynamic weighting mechanisms, which restricts comprehensive health assessment. To address these issues, this work fuses WiFi communication information with multi-modal wearable sensing data to develop an intelligent monitoring system featuring dynamic threshold adjustment and multi-path fault-tolerant decision-making. This study aims to jointly model physiological, motion, and environmental data, while improving real-time performance and computational efficiency through on-device data fusion and lightweight edge inference. It is hypothesized that multi-modal feature fusion and dynamic alert mechanisms can significantly enhance detection accuracy and stability, while WiFi optimization and compression-compensation mechanisms can effectively reduce latency and strengthen transmission robustness. Under low-latency and high-reliability conditions, the proposed system demonstrates adaptability to complex environments, providing a scalable and highly innovative solution for smart elderly care and remote healthcare applications.

2 Materials and methods

2.1 Construction of wearable device monitoring system based on WiFi communication

To achieve real-time monitoring and efficient transmission of multidimensional health status of the elderly, a multi-modal wearable elderly monitoring model based on WiFi communication technology is developed. The model integrates functional modules such as physiological signal acquisition, environmental parameter monitoring, WiFi data transmission, and cloud analysis, and achieves low latency and high reliability data processing and abnormal alarm through cloud-edge-end collaborative architecture. Wearable terminals are the core of data acquisition and preliminary processing, and their hardware module composition is shown in Figure 1.

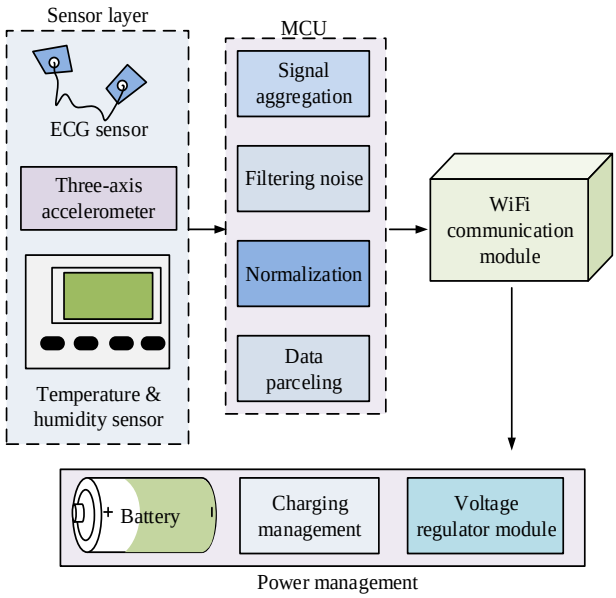


Figure 1: Hardware module composition of the wearable terminal

As shown in Figure 1, wearable terminals are composed of multiple types of sensors, Microcontroller Units (MCUs), power management modules, and WiFi communication modules. Electrocardiogram (ECG) sensors can continuously acquire ECG signals, three-axis accelerometers are used to obtain human motion status information, and temperature and humidity sensors monitor living environment parameters [11-12]. The MCU is responsible for aggregating and preprocessing multi-source signals, and real-time communication with edge nodes through low-power WiFi modules [13]. The power management module ensures stable operation of the device under long-term wearing conditions, improving the comfort and reliability of daily use for the elderly. In the signal acquisition stage, the ECG signal first passes through a bandpass filter to remove power frequency interference and baseline drift, as shown in equation (1) [14].

$$S_{ECG}(t) = F_{bp}(x(t), f_1, f_h) \quad (1)$$

In equation (1), $x(t)$ represents the original ECG signal. $S_{ECG}(t)$ represents the signal processed by bandpass filtering. t represents the sampling time of the signal. F_{bp} represents the bandpass filter operator. f_1 and f_h represent the low-frequency and high-frequency cutoff frequencies of the filter, respectively. To suppress baseline drift and high-frequency noise, a fourth-order Butterworth IIR bandpass filter with cut-off frequencies of 0.5 Hz and 50 Hz is applied, which effectively retains the primary physiological components such as QRS complexes. The filtering is implemented using the SciPy signal processing library in Python, with mirror padding at signal boundaries to minimize phase distortion. The composite acceleration value of three-axis acceleration in motion state assessment is shown in equation (2).

$$A(t) = \sqrt{a_x^2(t) + a_y^2(t) + a_z^2(t)} \quad (2)$$

In equation (2), $a_x(t)$, $a_y(t)$, and $a_z(t)$ represent the acceleration components in the three directions of x, y, and z, respectively. $A(t)$ represents the synthetic acceleration, used for fall detection and activity pattern recognition. Environmental parameters need to be normalized before transmission to ensure comparability of features with different dimensions. The normalization is shown in equation (3).

$$E_{norm}(i) = \frac{E_i - E_{min}}{E_{max} - E_{min}} \quad (3)$$

In equation (3), E_i represents the environmental parameter value of the i-th sampling point. E_{min} and E_{max} represent the minimum and maximum values of the parameter, respectively. $E_{norm}(i)$ represents the normalized environmental parameter value. After completing the collection and preprocessing, the data is transmitted to the edge or cloud for analysis through a WiFi link, as shown in Figure 2.

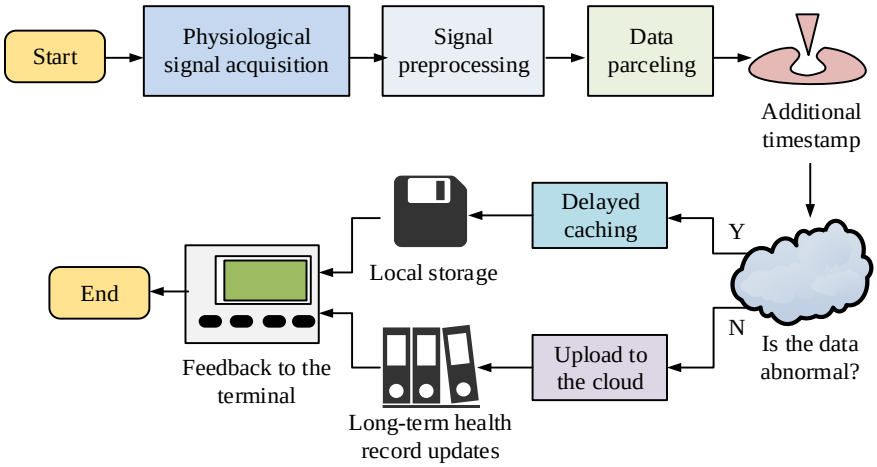


Figure 2: Data acquisition and WiFi transmission process

As shown in Figure 2, the entire data collection and WiFi transmission process is carried out in chronological order. Firstly, wearable terminals collect multi-source signals and perform preprocessing operations such as filtering, denoising, and normalization on the original signals to ensure the stability and consistency of the input data. Secondly, the MCU divides the preprocessed data into packets and adds timestamps, and sends the data to the edge nodes through the WiFi module. Edge nodes quickly analyze the received data to determine if there are any abnormal events. When the detection result is abnormal, the data will be prioritized for reporting to the cloud for in-depth analysis. When the detection result is normal, the data will enter the latency cache or be synchronized to the cloud according to the set cycle to reduce network bandwidth usage [15]. Finally, after completing the multi-modal fusion analysis, the cloud will return the monitoring results and alarm information to the terminal, achieving local prompts or emergency alarms. The WiFi link quality can be estimated by the received signal strength indicator to determine the transmission distance, as shown in equation (4).

$$d = 10^{\frac{P_t - P_r - L_0}{10n}} \quad (4)$$

In equation (4), P_t and P_r represent the signal strength at the transmitting and receiving ends, respectively. L_0 represents the reference distance path loss. n represents the path loss index. d represents the estimated transmission distance. In anomaly detection, an adjustable hyperparameter α_1 is introduced to set a dynamic warning threshold, as shown in equation (5).

$$\delta(t) = \begin{cases} 1, V(t) \geq \alpha \cdot \mu V \\ 0, \text{otherwise} \end{cases} \quad (5)$$

In equation (5), $V(t)$ represents the monitoring value

of a single tube. μV represents the mean under normal conditions. α represents the adjustment coefficient of the warning threshold. $\delta(t)$ represents the warning signal flag. When the value is 1, it indicates triggering the warning. When the value is 0, it indicates not triggering. The total latency of the system is composed of collection latency, transmission latency, and processing latency, as shown in equation (6).

$$T_{total} = T_{acq} + T_{tx} + T_{proc} \quad (6)$$

In equation (6), T_{acq} represents the acquisition latency. T_{tx} represents WiFi transmission latency. T_{proc} represents the processing latency at the edge or in the cloud. T_{total} represents the overall latency of the system. After combining hardware modules, data collection and transmission processes, and cloud-edge-end collaborative mechanisms, a complete overall architecture is formed, as shown in Figure 3.

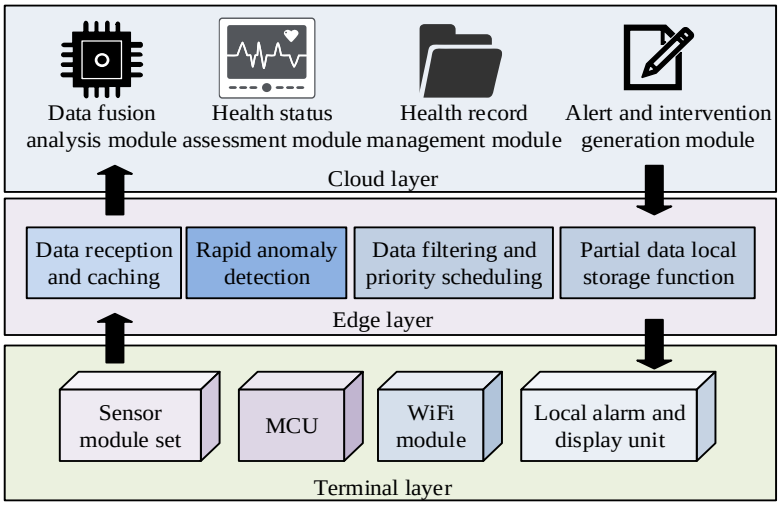


Figure 3: Overall architecture of the wearable elderly monitoring

As shown in Figure 3, the overall architecture of the monitoring system is divided into three parts: terminal layer, edge layer, and cloud layer. The terminal layer is responsible for real-time data collection and preliminary processing, including multiple types of sensor modules, data processing units, and WiFi communication modules. It is also equipped with local alarm and display functions to directly alert users in emergency situations. The edge layer is responsible for data reception and rapid screening, and can immediately perform lightweight anomaly detection upon receiving data, and prioritize scheduling based on the importance and urgency of the data. The cloud layer performs multi-modal data fusion analysis and long-term health record management, utilizing the combination of historical and real-time data to achieve accurate health status assessment and prediction, and generating personalized warning and intervention plans. The results are then fed back to the edge and terminal layers to ensure that the monitoring system can maintain stable and efficient operation under different network conditions.

2.2 Performance optimization of wearable monitoring system based on multi-modal fusion and dynamic weighting

After constructing a wearable device monitoring system based on WiFi communication, to further improve the data transmission stability and anomaly detection accuracy in complex home environments, the research introduces data compression and batch transmission mechanisms, multi-modal fusion algorithms, and dynamic threshold and weight adjustment methods [16-17]. The basic framework of this optimization method is shown in Figure 4.

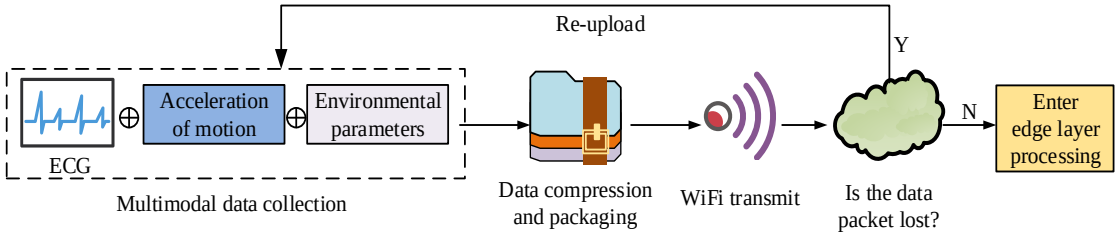


Figure 4: WiFi Transmission optimization and packet loss compensation flow

As shown in Figure 4, the WiFi transmission optimization and packet loss compensation flow introduces several improvements to the original system. Firstly, after completing data acquisition and preprocessing, the terminal compresses and batches the data to reduce transmission frequency and network load. During WiFi transmission, the system continuously monitors link quality and packet loss, automatically triggering re-transmission for missing or corrupted packets to ensure data integrity. When receiving data, the edge node prioritizes abnormal events and dynamically adjusts the alert threshold and feature weighting according to variations in real-time Signal-to-Noise Ratio (SNR), signal fluctuation amplitude, and packet loss rate. This adaptive mechanism allows the system to maintain detection stability and sensitivity under changing environmental conditions. When signal quality degrades or noise increases, the threshold is raised to suppress false alarms. Under stable conditions, it is lowered to enhance responsiveness. The adjustment logic operates in real-time at the edge node using a sliding-window strategy without adding computational overhead. Through these optimizations in transmission strategy and fault tolerance, the system effectively reduces latency and network pressure while enhancing stability and reliability in complex environments. In the optimized data transmission and compensation stage, the Signal Magnitude Vector (SMV) is introduced to calculate the feature extraction of three-axis acceleration. The calculation is shown in equation (7) [18].

$$\mathrm{SMV}(t)=\sqrt{(a_x(t)-\bar{a}_x)^2+(a_y(t)-\bar{a}_y)^2+(a_z(t)-\bar{a}_z)^2} \tag{7}$$

In equation (7), $\bar{a}_x$, $\bar{a}_y$, and $\bar{a}_z$ respectively represent the sliding average of axial acceleration, which is used to eliminate long-term bias and enhance sensitivity and robustness to fall events. Meanwhile, to monitor heart rate abnormalities, Heart Rate Variability (HRV) is introduced as a feature, as shown in equation (8) [19].

$$\mathrm{HRV}=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RR_i-\overline{RR})^2} \tag{8}$$

In equation (8), RR_i represents the i-th heartbeat interval. $\overline{RR}$ represents the average interval between heartbeats. N represents the total number of heartbeats. HRV features are used together with SMV as inputs for fall detection and anomaly determination modules at edge nodes, ensuring the integrity of data transmission information. Based on the calculated SMV and HRV features, they can be further fed into the fall detection algorithm for classification. The overall process is shown in Figure 5.

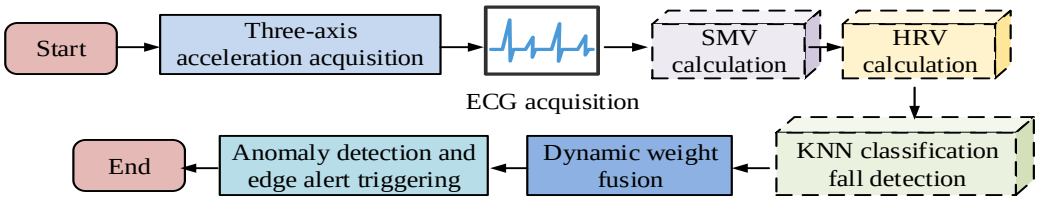


Figure 5: Fall detection algorithm flow

As shown in Figure 5, the fall detection algorithm process is designed based on multi-modal feature fusion. The algorithm first collects acceleration and heart rate signals through the terminal, and calculates synthetic motion features and HRV. Subsequently, the processed features are input into the classifier for fall detection, and the detection result triggers an abnormal alarm signal. Edge nodes decide whether to immediately report to the cloud for further analysis or cache locally for synchronization based on alarm priority. This algorithm improves the accuracy and robustness of fall event recognition by integrating motion and physiological signal features. In this process, the K-Nearest Neighbors (KNN) classifier is employed at the edge node for multi-modal feature classification. Compared with deep or kernel-based methods that require extensive parameter training and structural optimization, KNN offers low computational cost, no need for model pretraining, and easy deployment on resource-constrained devices. It ensures reliable classification accuracy while significantly reducing latency and energy consumption, aligning well with the real-time requirements of the proposed cloud-edge-end collaborative architecture. The specific implementation method is shown in equation (9) [20].

$$y=\arg\max_{c\in C}\sum_{i\in K}w_i\cdot\theta(y_i=c) \tag{9}$$

In equation (9), C represents the set of categories. K represents the set of nearest neighbor samples. w_i represents the weight of each neighbor. θ represents the indicator function. y_i represents the category of neighbors. To quantify the optimization effect of WiFi transmission, the system evaluates the link quality through a Packet Loss Rate (PLR), as shown in equation (10).

$$P_{loss}=\frac{N_{lost}}{N_{total}} \tag{10}$$

In equation (10), N_{lost} represents the number of lost packets. N_{total} represents the total number of data packets sent. P_{loss} represents the PLR. While optimizing transmission efficiency, the data compression ratio is used to measure the compression effect, as shown in equation (11).

$$R_{comp} = \frac{S_{original} - S_{compressed}}{S_{original}} \quad (11)$$

In equation (11), $S_{original}$ represents the size of the original data. $S_{compressed}$ represents the compressed data size. R_{comp} represents the compression ratio. The improved compression ratio not only reduces the network load, but also reduces the terminal energy consumption, providing more efficient data support for subsequent edge computing and cloud analysis. The multi-modal features are fused at the edge node to form the final decision output, and the weight fusion is shown in equation (12).

$$V_{fused} = \beta \cdot V_{SMV} + (1-\beta) \cdot V_{HRV} \quad (12)$$

In equation (12), V_{SMV} represents the motion feature value. V_{HRV} represents the HRV feature value. β represents the weight adjustment coefficient. V_{fused} represents the output after fusion. When designing the multi-modal fusion module, a linear weighted model is selected due to its simplicity, interpretability, and low computational cost, making it well suited for real-time inference on wearable and edge devices. Compared with more complex fusion strategies such as attention mechanisms or ensemble learning, the linear model offers higher energy efficiency and easier deployment without additional training or memory overhead. Subsequent experimental results further demonstrate that this choice provides a balanced trade-off between detection accuracy and latency. After calculating the weight fusion of each algorithm, the system integrates multi-modal features with optimization strategies to form a complete monitoring process, as shown in Figure 6.

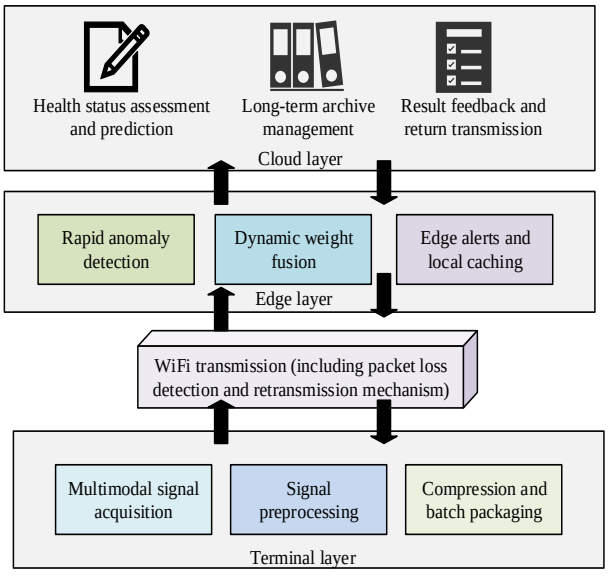


Figure 6: Optimized complete system workflow

As shown in Figure 6, the optimized complete system workflow integrates the end, edge, and cloud layers together with various optimization strategies. Firstly, the terminal performs multi-modal data acquisition and preprocessing, extracting key physiological and environmental features. The MCU then segments the data, appends timestamps, and transmits it through the optimized WiFi process while applying packet loss compensation, data compression, and batch transmission mechanisms. The edge node conducts rapid anomaly detection and performs multi-path fusion-based consistency analysis to handle potential signal conflicts. When discrepancies arise between different modalities (e.g., physiological and motion features), the system re-evaluates them based on confidence levels and temporal variation trends. If conflicting signals persist and the overall anomaly confidence increases, a prioritized alert is triggered and forwarded to the cloud for further verification. If the inconsistency is identified as transient noise or posture variation, the alert is temporarily suppressed to avoid false alarms. Finally, the cloud performs deep fusion analysis and health status evaluation, returning monitoring results and alerts to the terminal to achieve a closed-loop monitoring process. This workflow demonstrates the comprehensive optimization of the system in multi-mode conflict handling, low latency, high reliability, and high accuracy. To clarify the computational responsibilities across different layers, the proposed system defines a clear split between the terminal, edge, and cloud components. The terminal layer handles multi-modal data acquisition and basic preprocessing, including ECG filtering, motion sensing, and RSSI monitoring. The edge layer performs HRV feature computation, feature normalization, and KNN-based anomaly detection for real-time response. The cloud layer manages large-scale data fusion, model training, and adaptive system optimization. The functional division of tasks is summarized in Table 1.

Table 1: Functional division of tasks across terminal, edge, and cloud layers

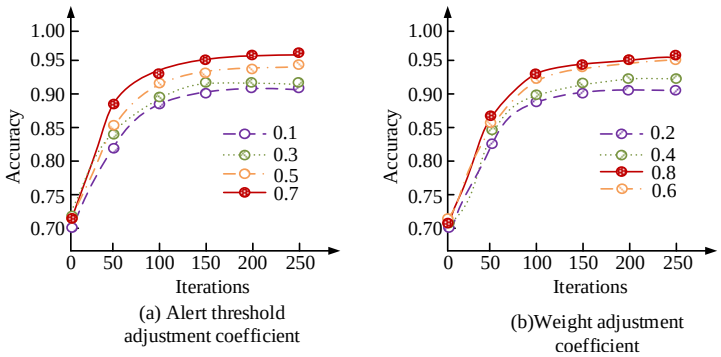
Layer	Main tasks	Module	Output
End	Data acquisition & preprocessing	AD8232 ECG, MPU-6050 motion sensing, RSSI monitoring	Filtered signals, motion data
Edge	Feature computation & local classification	HRV calculation, feature normalization, KNN classifier	Real-time alerts
Cloud	Global analysis & adaptive optimization	Deep fusion model training, user profiling, system tuning	Long-term assessment & model updates

3 Results and analysis

3.1 Hyperparameter selection and performance ablation testing for wearable monitoring systems

During the experimental phase, a hardware and software operating environment suitable for joint monitoring of wearable devices and WiFi communication is established. The CPU is Intel Xeon Gold 6338, the GPU is NVIDIA RTX A6000 48GB, and the motherboard supports Gigabit Ethernet and WiFi 6 wireless networks, ensuring high-speed data exchange between the end, edge, and cloud. The software runs on the Ubuntu 20.04 LTS operating system, with Python 3.10 and TensorFlow 2.12 deep learning frameworks deployed on edge nodes. A Flask-based data interface is configured in the cloud to collect, transmit, analyze, and store monitoring data. In addition, the wearable terminal incorporates an AD8232 analog front-end ECG module for 0.5-50 Hz cardiac signal acquisition and an MPU-6050 accelerometer-gyroscope sensor for motion and posture tracking. Wireless communication relies on ESP32 microcontrollers that support WiFi 6. The WiFi signal quality is evaluated using the Received Signal Strength Indicator (RSSI), with a time-weighted averaging scheme applied to smooth instantaneous fluctuations and reflect link stability under varying interference conditions. The power consumption model is derived from each module's rated power and duty cycle, with total energy consumption expressed in milliwatts (mW) to evaluate energy distribution and efficiency under different operational states.

To verify the performance of the system in fall detection, health status recognition, and multi-modal fusion judgment, the study selects the SisFall Falls and Daily Activities Dataset (SisFall) and the MIT-BIH Physiological Signal Database (MIT-BIH) as the test datasets. The SisFall Dataset and the MIT-BIH Database are selected as the testing datasets. The SisFall dataset, jointly developed by Universidad de los Andes (Colombia) and Universidad Técnica Federico Santa María (Chile), contains tri-axial acceleration and angular velocity data collected from subjects wearing accelerometers and gyroscopes performing 15 fall scenarios and 19 daily activities. The participants are mainly middle-aged and elderly, providing representative motion features and activity patterns suitable for validating wearable fall detection algorithms. The MIT-BIH database includes ECG, blood pressure, and various physiological signals from participants across different age ranges. It has been widely used for HRV analysis and arrhythmia detection, supporting the evaluation of system stability and generalization under diverse physiological conditions. During data preprocessing, ECG signals are filtered using a 0.5-50 Hz bandpass filter to remove baseline drift and high-frequency noise. The sampling rate is unified at 360 Hz, and HRV features are extracted using a 10-second sliding window with 50% overlap. The tri-axial acceleration signals in the SisFall dataset are sampled at 200 Hz and processed through normalization and temporal alignment to ensure consistency across subjects and synchronization with ECG data. To determine the optimal performance parameter configuration for each module in the system, a hyperparameter selection test is conducted on the warning threshold adjustment coefficient α and weight adjustment coefficient β. The test results are shown in Figure 7.



(a) Alert threshold adjustment coefficient

(b)Weight adjustment coefficient

Figure 7: Hyperparameter test results

Figure 7 (a) shows the results of the selection test for the warning threshold adjustment coefficient α, and Figure 7 (b) shows the results of the selection test for the weight adjustment coefficient β. As shown in Figure 7 (a), with the gradual increase of α, the overall accuracy of the system during the training process showed an upward trend. When α=0.7, the proposed system could quickly improve accuracy at multiple iteration points and reach the highest value of 96.4% when the iteration approached the stable stage. When α was set to 0.3 or below, the accuracy improvement was slow and the final level was low, indicating that a small threshold adjustment was not sufficient to effectively determine abnormal events. As shown in Figure 7 (b), with the increase of β, the accuracy of the system in multi-modal feature determination gradually improved. When β=0.6, the accuracy reached the highest 95.8%, indicating that the fusion weight allocation between motion features and HRV features was most reasonable at this time, which could fully utilize the complementary advantages of the two types of information and improve the recognition accuracy of falls and health abnormalities. Therefore, the study ultimately determined the optimal combination of hyperparameters as α=0.7 and β=0.6, under which the overall performance of the system was optimal. The study continues to conduct ablation tests on each module of the system on two datasets, and the results are shown in Figure 8.

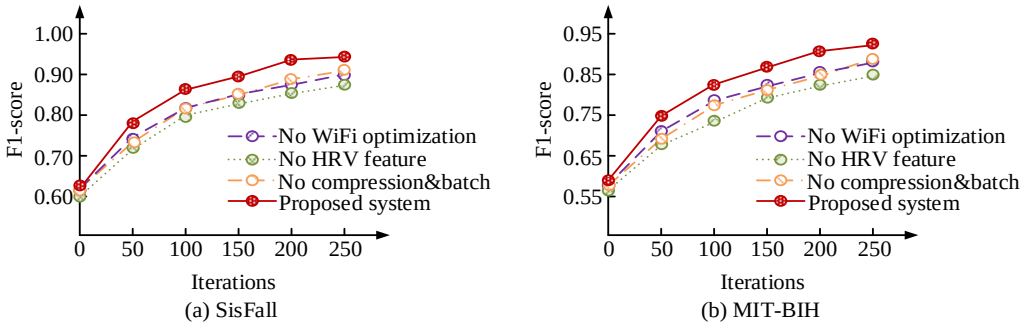


Figure 8: Ablation test results

Figure 8 (a) shows the ablation test results on the SisFall dataset, and Figure 8 (b) shows the results on the MIT-BIH dataset. According to Figure 8 (a), the F1-score of the complete system stabilized at 0.94 after 250 iterations. In contrast, after removing WiFi transmission optimization, the final F1-score was 0.90. After removing HRV features, it decreased to 0.88. After removing data compression and batch transmission, it was 0.91. Similarly, in Figure 8 (b), the F1-score of the complete system ultimately reached 0.92, higher than that of the other three baseline models. The ablation test results indicate that deleting all three modules will lead to a decrease in F1-score. The ECG scene is more sensitive to HRV features, and removing the HRV feature module has the greatest impact on the overall performance of the system. To evaluate the impact of each functional module on the overall system performance, comparative experiments were conducted on the SisFall and MIT-BIH datasets. The results are summarized in Table 2.

Table 2: Test results for different indicators on SisFall and MIT-BIH datasets

Data set	Method	Precision/%	Recall/%	Specificity/%
SisFall	No WiFi optimization	94.5	92.1	95.0
	No HRV feature	93.2	90.6	94.0
	No compression&batch	93.9	91.3	94.7
	Proposed system	97.0	95.4	97.6
MIT-BIH	No WiFi optimization	94.2	91.9	94.0
	No HRV feature	92.8	90.2	92.5
	No compression&batch	93.5	91.0	93.3
	Proposed system	96.1	94.8	96.3

According to Table 2, on the SisFall dataset, the precision of the complete system reached 97.0%, the recall rate was 95.4%, and the specificity was 97.6%. On the MIT-BIH dataset, the system also maintained a leading position with a precision of 97.6%, a recall rate of 94.8%, and a specificity of 96.3%. In contrast, the version without HRV features showed a significant decrease in recall rates on both datasets, at 90.6% and 90.2%, respectively, indicating that this feature played a crucial role in identifying falls and heart rate abnormalities. After removing WiFi optimization or data compression and batch transmission, although the three

indicators slightly decreased, the overall performance still exceeded 91%, verifying that the system had robustness.

3.2 Simulation testing of wearable monitoring system

To further verify the adaptability and stability of the proposed wearable elderly monitoring system under different practical scenarios, three types of simulation environments, including low, medium, and high interference, were constructed based on the SisFall and MIT-BIH real-world datasets, corresponding to daily home, outdoor, and complex field monitoring conditions. To create representative simulation environments, multiple interference factors were introduced into the original datasets, including sensor noise, motion artifacts, and WiFi link attenuation. Specifically, zero-mean Gaussian white noise was added to the acceleration and ECG signals to simulate sampling errors and ambient electromagnetic interference, while amplitude fluctuations and periodic disturbances were introduced to reproduce body motion and posture variations during device usage. In addition, the SNR and path loss exponent (n) were adjusted to emulate network instability during WiFi data transmission. To ensure reproducibility, the interference intensity was classified according to noise variance and SNR parameters: In the low-interference environment, the noise variance was set to 0.01 with SNR above 30 dB; In the medium-interference environment, the noise variance was 0.05 with SNR around 20 dB; In the high-interference environment, the noise variance was 0.1 with SNR reduced to 10 dB, accompanied by randomly injected motion artifact signals. These interferences were applied to the original samples through weighted addition during the data preprocessing stage to maintain the structural and temporal consistency of the signals.

In the testing, three widely used comparison systems are selected, namely the Hidden Markov Model-based Fall Detection (HMM-FD), Convolutional LSTM-based Wearable Monitoring (ConvLSTM-WM), and Graph Convolutional Network-based Pose Monitoring (GCN-Pose). The study uses sensitivity as an indicator to test the response capability of each system to abnormal events in three types of simulation environments. The test results are shown in Figure 9.

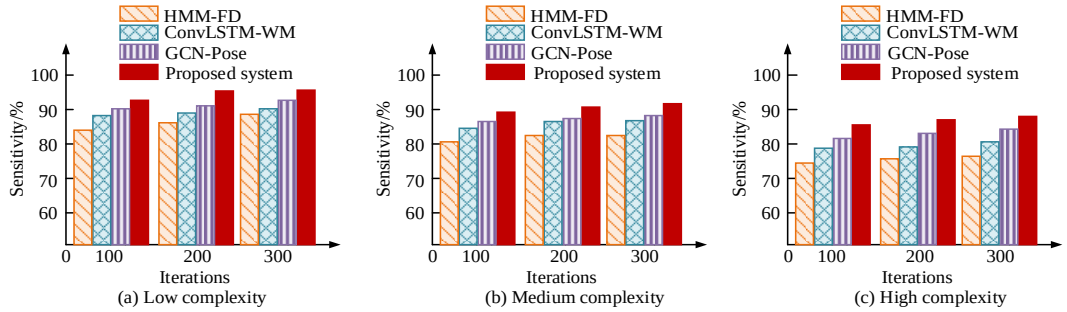


Figure 9: Sensitivity comparison of four wearable monitoring systems

Figures 9 (a), 9 (b), and 9 (c) respectively show the sensitivity changes of four systems in low, medium, and high complexity simulation environments. In Figure 9 (a), the proposed system increased from 92.8% to 94.8% in low complexity, which was 2.9%, 4.6%, and 7.7% higher than that of GCN-Pose, ConvLSTM-WM, and HMM-FD, respectively. At the medium complexity in Figure 9 (b), the sensitivity of the proposed system increased from 89.5% to 91.4%, while GCN-Pose, ConvLSTM-WM, and HMM-FD were 87.5%, 85.8%, and 82.3%, respectively. In the high complexity scenario shown in Figure 9 (c), the overall sensitivity of each system decreased, while the proposed system increased from 85.0% to 86.8%, still leading the three comparison systems. Overall, the proposed system maintains the highest sensitivity in each type of simulation environment, indicating that the system's ability to detect abnormal events is more robust in simulating real home interference and multi-action scenarios.

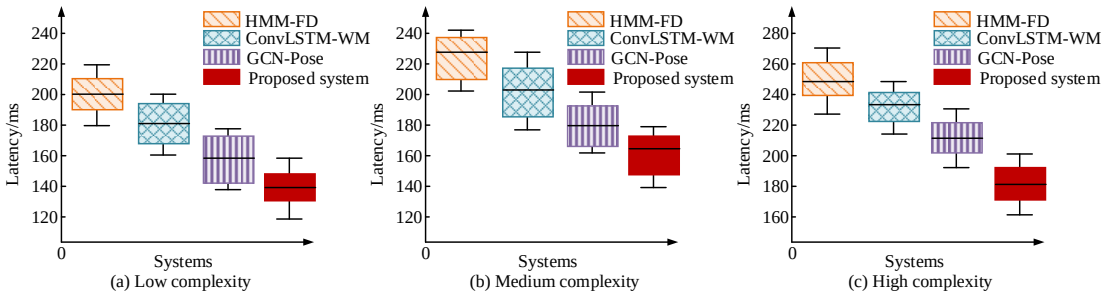


Figure 10: Event detection latency results

Figures 10 (a), 10 (b), and 10 (c) respectively show the test results of event detection latency box plots for four systems under three different levels of complexity. There were significant differences in latency among systems in low complexity, medium complexity, and high complexity environments. As shown in Figure 10 (a), in the low complexity, the latency of HMM-FD was relatively high, with a median of 200 milliseconds, while the median of the proposed system was only 140 milliseconds. From Figure 10 (b), in the medium complexity environment, the latency of all four systems increased, but the proposed system still maintained the lowest median latency of 160 milliseconds, which was better than that of other systems. In Figure 10 (c), high complexity testing resulted in HMM-FD latency reaching 250 milliseconds, while the median latency of the proposed system was 180 milliseconds, maintaining optimal performance. Overall, the proposed system exhibits low latency and high stability in all three types of simulation environments, proving its excellent event response capability and robustness under different levels of complexity and signal interference conditions. To clearly illustrate the overall performance of each monitoring system under different simulation environments, three key metrics, sensitivity, F1-score, and detection latency were summarized and compared. The results are shown in Table 3.

Table 3: Test results of classification performance under different conditions

Situation	Method	F1-score	Sensitivity/%	Latency /ms
Low complexity	HMM-FD	0.89	87.1	200
	ConvLSTM-WM	0.91	90.2	175
	GCN-Pose	0.92	91.9	165
	Proposed system	0.94	94.8	140
Medium complexity	HMM-FD	0.86	82.3	225
	ConvLSTM-WM	0.88	85.8	198
	GCN-Pose	0.89	87.5	188
	Proposed system	0.91	91.4	160
High complexity	HMM-FD	0.83	79.4	250
	ConvLSTM-WM	0.85	82.1	230
	GCN-Pose	0.86	83.6	215
	Proposed system	0.88	86.8	180

As shown in Table 3, the proposed system achieved F1-scores of 0.94, 0.91, and 0.88 under low, medium, and high interference environments, respectively, higher than that of HMM-FD, ConvLSTM-WM, and GCN-Pose, indicating superior detection accuracy and stability across different scenarios. Considering the overall performance in sensitivity and latency, the proposed system consistently balanced detection accuracy and response speed under varying complexity, demonstrating strong robustness and adaptability. Statistical significance analysis further confirmed that the improvements in F1-score (p=0.032) and detection latency (p=0.019) were significant at the α=0.05 level, indicating that the performance advantage was statistically meaningful rather than random variation.

The study continues to compare the test results of four models under three different operating conditions, including False Alarm Rate (FAR), Energy Consumption (EC), and PLR. When calculating EC, direct hardware measurements were not performed. On the contrary, these values were estimated based on the typical power consumption parameters of key wearable terminal components, including the MCU, sensors, and WiFi communication module. Specifically, the average power demand of each component was derived from its rated operating voltage and average current under different activity states, and the overall energy consumption was obtained by combining these values with the system's task execution time. All parameters were taken from publicly available component datasheets and experimental runtime logs, reliably estimating the system's energy consumption trends during the cloud-edge-end collaborative operation. The results are shown in Table 4.

Table 4: Test results under different conditions

Situation	Method	FAR%	EC/mW	PLR/%
Low complexity	HMM-FD	7.5	82.3	1.8
	ConvLSTM-WM	6.2	95.1	1.5
	GCN-Pose	5.5	110.4	1.2
	Proposed system	3.8	88.7	0.9
Medium complexity	HMM-FD	9.1	84.5	2.6
	ConvLSTM-WM	7.3	97.8	2.1
	GCN-Pose	6.5	112.0	1.8
	Proposed system	4.5	90.3	1.1
High complexity	HMM-FD	11.4	87.2	3.5
	ConvLSTM-WM	9.0	101.2	2.9
	GCN-Pose	7.8	115.5	2.3
	Proposed system	5.2	92.1	1.4

According to Table 4, in the low complexity environment, the HMM-FD system had the weakest performance with a FAR of 7.5%, EC of 82.3mW, and signal PLR of 1.8%. In contrast, the proposed system had a FAR of only 3.8%, EC of 88.7mW, and signal PLR of 0.9% in low complexity environments, demonstrating high efficiency and reliability. In the medium complexity environment, the performance gap between systems further widened, and the proposed system still maintained a low FAR of 4.5% in this environment. In the high complexity environment, the FARs of HMM-FD and ConvLSTM-WM increased to 11.4% and 9.0%, respectively, and the signal PLR increased significantly. The proposed system still maintained the lowest FAR of 5.2%, signal PLR of 1.4%, and EC of 92.1mW, proving that the system exhibited good comprehensive performance and strong robustness in accuracy, reliability, and EC control.

4 Discussion

4.1 Privacy and security considerations

Given that the proposed elderly monitoring system involves the acquisition and transmission of physiological and behavioral data, data privacy and security have been carefully considered in the system design. Communication between the end and edge nodes employs AES-256 encryption to prevent data interception or tampering during WiFi transmission. Cloud-side data storage and access are protected through user authentication and permission control to avoid unauthorized access. While the current study primarily focuses on monitoring and detection performance, future work will incorporate end-to-end encryption, data anonymization, and compliance with international privacy standards such as GDPR, further enhancing system reliability and security in healthcare applications.

4.2 System performance analysis and comparative discussion

This study develops a multi-mode monitoring system that integrates WiFi communication and wearable sensing, which solves the challenges of unstable signals, large movement changes, and latency in elderly health monitoring. The system is systematically validated on two public datasets, SisFall and MIT-BIH. The experimental results show that the proposed system outperforms HMM-FD, ConvLSTM-WM, and GCN-Pose in F1-score, sensitivity, and detection latency, demonstrating strong robustness and real-time responsiveness. The performance improvement mainly stems from three aspects. Firstly, the HRV analysis at the feature level allows the system to capture both motion and physiological anomalies, improving the accuracy of abnormal event detection. Secondly, the WiFi transmission optimization and compression-compensation mechanism effectively reduces data loss and latency, ensuring stability under complex interference conditions. Thirdly, the edge-cloud collaborative architecture with dynamic threshold adjustment achieves balanced computation and communication loads, enhancing performance under high concurrency.

Compared with existing approaches, HMM-FD relies on statistical modeling and struggles with multi-modal data. ConvLSTM-WM achieves strong temporal modeling but suffers from high computational cost and energy consumption. GCN-Pose performs well in pose estimation but is more sensitive to interference. In contrast, the proposed system jointly models physiological and motion features, achieving a superior trade-off between detection accuracy and power efficiency.

Nevertheless, certain limitations remain. The dynamic adjustment mechanism may increase energy consumption in multi-node scenarios, and system stability can still be affected by WiFi coverage and environmental occlusion. Future work will explore federated learning and adaptive energy management strategies to improve distributed performance, privacy protection, and long-term operational reliability.

5 Conclusions and recommendations

A monitoring system combining WiFi transmission optimization, HRV feature analysis, and multi-modal data fusion was proposed to address the data noise interference, motion complexity, and multi-modal feature fusion in fall detection and health status monitoring of wearable elderly monitoring systems. The system rapidly identified abnormal events by adjusting the warning threshold coefficient and optimizing the fusion weight, and combined the end-to-end cloud collaborative architecture to ensure efficient data transmission and processing. The hyperparameter tests on the SisFall and MIT-BIH real datasets showed that when the warning threshold adjustment coefficient was set to 0.7 and the fusion weight coefficient was 0.6, the system performance was optimal, with F1-scores reaching 0.94 and 0.92, respectively. In a simulation environment, the proposed system was compared with three typical monitoring systems. The sensitivity of the proposed system could reach up to 94.8% in low, medium, and high interference environments, with a minimum event detection latency of 140 milliseconds. The minimum FAR of the system was only 3.8%, and the signal PLR was 0.9%, fully reflecting its strong robustness and real-time response capability under complex actions and interference conditions.

In summary, the WiFi-based wearable monitoring scheme demonstrates unique advantages in non-contact sensing, low-power operation, and environmental adaptability. Compared with traditional systems that rely on single-sensor input or control-oriented feedback, it exhibits higher reliability and deployment flexibility. Future research may further integrate artificial intelligence techniques to enable health risk prediction and trend analysis, utilizing temporal deep learning and graph neural network models to enhance health status assessment and early warning capabilities. In addition, research focused on battery optimization, multi-node collaborative scalability, and user-friendly interaction design can further promote the long-term application and large-scale deployment of wearable monitoring systems in intelligent elderly care and home healthcare scenarios.

References

[1] Chen C, Ding S, Wang J. Digital health for aging populations. Nature medicine, 2023, 29(7): 1623-1630. DOI: 10.1038/s41591-023-02391-8.

[2] Ledesma J R, Ma J, Zhang M. Global, regional, and national age-specific progress towards the 2020 milestones of the WHO End TB Strategy: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet infectious diseases, 2024, 24(7): 698-725. DOI: 10.1016/S1473-3099(24)00007-0.

[3] Naeini E K, Sarhaddi F, Azimi I. A deep learning–based PPG quality assessment approach for heart rate and heart rate variability. ACM Transactions on Computing for Healthcare, 2023, 4(4): 1-22. DOI: 10.1145/3616019

[4] Babu M, Lautman Z, Lin X. Wearable devices: implications for precision medicine and the future of health care. Annual review of medicine, 2024, 75(1): 401-415.DOI: 10.1146/annurev-med-052422-020437.

[5] Dewangan L, Vaishnava A R. Energy-efficient smart wearable IoT device for the application of collapse motion detection and alert. IETE Journal of Research, 2023, 69(2): 1133-1139. DOI: 10.1080/03772063.2020.1859952.

[6] Tan L, Yu K, Bashir A K, Cheng X, Ming F, Zhao L, Zhou X. Toward real-time and efficient cardiovascular monitoring for COVID-19 patients by 5G-enabled wearable medical devices: A deep learning approach. Neural Computing and Applications, 2023, 35(19): 13921-13934. DOI: 10.1007/s00521-021-06219-9.

[7] Chen J, Wang W, Fang B, Liu Y, Yu K, Leung V C M. Digital twin empowered wireless healthcare monitoring for smart home. IEEE Journal on Selected Areas in Communications, 2023, 41(11): 3662-3676. DOI: 10.1109/JSAC.2023.3310097.

[8] He J, Zhu W, Qiu L, Zhang Q, Wang C. An indoor fall detection system based on WiFi signals and genetic algorithm optimized random forest. Wireless Networks, 2024, 30(3): 1753-1771. DOI: 10.1007/s11276-023-03625-w.

[9] Ghoshal S, Saif S, Biswas S. Optimizing Indoor Positioning: ANN-GD Fusion for Enhanced Accuracy in WiFi Fingerprint-Based Surveillance Systems. SN Computer Science, 2025, 6(5): 511-512. DOI: 10.1007/s42979-025-04029-7.

[10] Zhang X, Gu Y, Yan H, Wang Y, Dong M, Ota K. Wital: A COTS WiFi devices based vital signs monitoring system using NLOS sensing model. IEEE Transactions on Human-Machine Systems, 2023, 53(3): 629-641. DOI: 10.1109/THMS.2023.3264247.

[11] Sirtoli V G, Liamini M, Lins L T. Removal of motion artifacts in capacitive electrocardiogram acquisition: A review. IEEE Transactions on Biomedical Circuits and Systems, 2023, 17(3): 394-412. DOI: 10.1109/TBCAS.2023.3270661.

[12] Ghasemi S, Sotoudeh B, Pazhooh M. Design and simulation of a novel MEMS-based single proof mass three-axis piezo-capacitive accelerometer. Microsystem Technologies, 2024, 30(3): 279-289. DOI: 10.1007/s00542-023-05604-9.

[13] Bellier P, Coppieters’t Wallant D, van den Bongarth H. A wireless low-power single-unit wearable system for continuous early warning score calculation. IEEE Sensors Journal, 2023, 23(11): 12171-12180. DOI: 10.1109/JSEN.2023.3267146.

[14] Gupta V. DBPF pre-processing-based improved ECG signal analysis in medical engineering applications. International Journal of Engineering Systems Modelling and Simulation, 2025, 16(2): 111-118. DOI: 10.1504/IJESMS.2025.144881.

[15] Lv Z, Singh A K. Edge-cloud-based wearable computing for automation empowered virtual rehabilitation. IEEE Transactions on Automation

Science and Engineering, 2023, 21(3): 3896-3909. DOI: 10.1109/TASE.2023.3289908.

[16] Reaño C, Riera J V, Romero V. A cloud-edge computing architecture for monitoring protective equipment. Journal of cloud computing, 2024, 13(1): 82-83. DOI: 10.1186/s13677-024-00649-1.

[17] Boulkroune A, Zouari F, Boubellouta A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, 2025: 10775463251320258.

[18] Renteria-Pinon M, Tang X, Tang W. Real-time in-sensor slope level-crossing sampling for key sampling points selection for wearable and iot devices. IEEE Sensors Journal, 2023, 23(6): 6233-6242. DOI: 10.1109/JSEN.2023.3243460.

[19] Lin I M, Chen T C, Tsai H Y. Four sessions of combining wearable devices and heart rate variability (HRV) biofeedback are needed to increase HRV indices and decrease breathing rates. Applied psychophysiology and biofeedback, 2023, 48(1): 83-95. DOI: 10.1007/s10484-022-09567-x.

[20] Kausar F, Mesbah M, Iqbal W. Fall detection in the elderly using different machine learning algorithms with optimal window size. Mobile Networks and Applications, 2024, 29(2): 413-423. DOI: 10.1007/s11036-023-02215-6.

FB2BPMN: An End-to-End Pipeline for Translating Unstructured User Feedback into BPMN Using LLM and Fuzzy Matching

Uce Indahyanti[1], Arif Djunaidy[2*], Daniel Siahaan[3]
[1,3]Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
[2]Department of Information Systems, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
E-mail: 7025211023@student.its.ac.id, arif.djunaidy@its.ac.id, daniel@if.its.ac.id
[*]Corresponding author

Keywords: BPMN, fuzzy string matching, NLP, LLM translation, user feedback text

Received: September 24, 2025

Translating unstructured user feedback into Business Process Model and Notation (BPMN) is challenging due to informal language, contextual ambiguity, and the lack of explicit structural cues. We present FB2BPMN, an end-to-end pipeline that combines natural language processing (NLP), large language models (LLMs), and fuzzy string matching to automatically generate BPMN elements from raw feedback. The pipeline comprises four stages: sentence structuring, fact extraction, role-activity mapping, and fuzzy-based semantic alignment. We evaluate FB2BPMN on 125 annotated feedback instances sampled from academic journal management systems. Using expert-authored BPMN as reference, FB2BPMN attains precision 0.97, recall 0.88, and F1 0.91 on element identification and accuracy 0.85 on process flow construction, outperforming a rule-based baseline. Results indicate strong structural and semantic correspondence, showing that FB2BPMN effectively bridges informal feedback and formal process representations.

Povzetek: Študija uvaja metodo FB2BPMN, ki samodejno generira BPMN modele iz ne-strukturiranih besedilnih povratnih informacij uporabnikov. Rezultati evalvacije na 125 primerih potrjujejo visoko natančnost in zanesljivost pristopa, kljub manjšim težavam pri poimenovanju elementov zaradi leksikalnih razli.

1 Introduction

Business Process Model and Notation (BPMN) is a widely adopted standard for graphically representing business processes in a domain-independent format understandable to diverse stakeholders [1], [2]. It bridges process design and implementation [3], enabling a shared understanding of system functionalities and operational flows [4]. Organizations typically maintain large collections of documents, ranging from formal operating procedures to unstructured textual feedback, that may contain process-relevant knowledge. While structured documentation facilitates BPMN generation in many real-world settings, formal sources are unavailable, and process knowledge exists primarily in unstructured form. BPMN models can be derived from various source documents, including structured sources such as business rules, standard operating procedures, and use case diagrams, as well as semi-structured or unstructured sources such as event logs, user stories, and user feedback [5]. Whereas structured sources enable straightforward BPMN generation, real-world user feedback is informal and ambiguous, making direct transformation more challenging. Generating BPMN from unstructured documents is particularly challenging due to inconsistent grammar, incomplete descriptions, and lack of formalized process elements, which are common issues when formal documentation is unavailable.

Prior research has explored SBVR-based approaches have been used to translate the interview transcripts into formal process models. BPMN generation from both structured and semi-structured sources. For example, SBVR-based approaches have been used to translate interview transcripts into formal process models [6], while user story-driven methods have been employed to construct BPMN diagrams through predefined semantic templates [7], [8]. These methods demonstrate high accuracy when processing well-formed input but struggle to handle the variability of informal, real-world feedback. User feedback can serve as a valuable asset for business process improvement. It can support incremental changes and explorative redesign initiatives [9], [10]. Today, the improvement or redesign of business process models still relies on expert expertise, often disregarding user feedback. Consequently, users often propose process redesigns, but these suggestions are frequently disregarded [11]. On the other hand, many users access app repositories and social media platforms to provide feedback on their digital experiences. This vast textual

corpus has the potential to be mined to extract detailed information regarding user expectations and feature-oriented suggestions [12].

Recent advances in natural language processing (NLP) and large language model (LLM) have opened new opportunities for generating process models directly from textual descriptions, including those derived from informal or unstructured sources such as user feedback, an area that remains underexplored. Fatawi et al. [13] conducted a bibliometric review underscoring the pivotal role of prompt engineering and LLM in NLP applications and their increasing relevance in automation systems. Studies have shown that LLM can extract structured information from natural language, map it to domain-specific ontologies [14], and support model-to-model transformation through semantic interpretation [15].

Recent studies highlight the potential of LLM in semantic understanding. Prasad et al. [16] demonstrated their effectiveness in interpreting unstructured text, supporting our use of LLM to translate user feedback into BPMN representations. Sonbol et al. [17] proposed a translation-inspired method to transform structured process descriptions into BPMN models, and the BPMN Chatbot tool [18] has demonstrated the ability to produce BPMN diagrams from text. Most existing methods for BPMN generation rely on structured inputs, such as formal descriptions or domain-specific templates. In contrast, using large language models to generate BPMN models from unstructured user feedback remains underexplored.

Research on business processes generally revolves around generating, redesigning, extracting, and measuring the similarity of business processes. Fehrer et al. [19] conducted process redesign and similarity measurement studies. The study proposed a redesign method utilizing the aBPR approach. Delias et al. [20] explored business process prototype improvements through evidence-based approaches, while [21] proposed a roadmap for process model improvement. Another study introduced a method for measuring business process similarity using role relation networks and role hierarchy graphs [22]. Furthermore, [23] present an approach to refine business process models by examining syntax and semantics. Work related to business process generation can be categorized into two main aspects: the source document type and the format of the resulting model. The following studies concentrate on building process models from structured text source documents, transforming them into formats or notations. Several researchers have proposed techniques for deriving business processes from source documents such as business rules and use cases, which are converted into SBVR [24], [25].

Other studies investigated the transformation of business rules and event logs into BPMN [11], [26], [27]. Methods for generating model processes from business rules and using case documents into SBVR were discussed [24]. SBVR translation into UML using an NLP approach was explored [25]. Other approaches transform textual business rules and event log documents into BPMN [11], [27]. Zeng, Duan et al. [28] describe top-down process mining as a method for generating process models from activity logs. Similarly, [29] propose an approach for constructing process models based on a set of refinement rules. In a related study, [30] investigate how Decision Model and Notation (DMN) can accurately represent data modeled within BPMN. Recent research on BPMN generation increasingly incorporates advanced technologies to improve automation and accuracy. Several studies [17], [31], [32] utilized textual requirements or textual description document sources. Sholiq et al. [33] demonstrated BPMN generation from Indonesian functional requirements using tailored syntactic patterns and parsing rules. These works combined machine learning techniques, including deep learning, with specific tools to improve the extraction and modeling processes from textual descriptions into BPMN diagrams. Textual business process documents can generate textual BPMN using a deep learning approach known as A-BPS [34].

Other approaches include a machine translation method to construct BPMN in a text-graph format [17] and the use of NLP and deep learning techniques, leveraging GPT language models and Graphviz tools [31]. Another method combines NLP techniques, SBVR definitions, rule sets, and spreadsheet-based descriptions to generate BPMN diagrams from textual requirements [32]. Danenas et al. [25] proposed an NLP-enhanced approach for extracting business vocabularies and rules from natural language, enabling formal representation aligned with SBVR standards. Tangkawarow et al. [6] proposed the ID2SBVR method to extract SBVR operational rules from informal document sources.

User feedback represents a valuable resource for improving business processes. Mustansir et al. [35] assessed user satisfaction by applying sentiment analysis to feedback related to process execution. In a subsequent study, Mustansir et al. [11] proposed classifying user feedback into three levels using deep learning and natural language processing techniques. Ahmed, Shahzad et al. [36] integrated user feedback into process models by analyzing relationships among feature vectors, data balancing strategies, and word embeddings in a different approach. A study in this domain presented a model-based approach for transforming user stories into process model notations [37]. Ahmed & Mustansir [38] introduced a novel NLP-based concept to map user feedback to relevant process model elements. An automated approach for generating business process models from user stories was proposed by [8]. Other research utilized user stories to construct business process models in UML format [7],[39]. Further studies focused on identifying and mapping user feedback into process elements [25], [31].

As discussed in the introduction, recent advances in NLP and LLM have opened opportunities for generating process models from text, particularly through tasks such as information extraction and semantic transformation [14],[15]. In the context of business process modeling, approaches such as the translation-inspired method by [17] and the BPMN Chatbot tool developed by [18] have demonstrated the feasibility of generating BPMN models from textual input. Conventional approaches to BPMN generation typically rely on structured inputs, limiting their adaptability to informal or ambiguous text. SBVR-

based methods require strict syntactic rules, making them sensitive to linguistic variability. User story techniques depend on predefined semantic templates, reducing flexibility in real-world scenarios. Deep learning models offer greater adaptability but often require extensive preprocessing and lack transparency. These constraints highlight the need for methods capable of handling unstructured inputs with minimal assumptions.

However, the automatic generation of BPMN from unstructured user feedback remains largely unexplored. Unlike user stories or formal requirements, real-world feedback often contains fragmented sentences, colloquial expressions, and multiple intertwined intents. Such characteristics make it difficult for template-based or rule-based approaches to produce accurate models. Existing LLM-based BPMN generation methods have primarily been tested on clean, syntactically consistent input, leaving a gap in applicability to noisy feedback data.

To our knowledge, no existing study has directly addressed the translation of unstructured user feedback into BPMN models. To fill this gap, we propose a method that converts informal feedback into BPMN representations without requiring prior structuring. The approach integrates natural language processing techniques, large language models, and fuzzy string matching to extract and map process elements from free-form text, enabling a more responsive and user-driven process modeling paradigm. This translation pipeline is designed to handle the variability and informality of user input, often containing fragmented expressions or symbols, and to produce structured textual representations suitable for BPMN model generation.

Several recent studies have explored BPMN generation from semi-structured text such as user stories, interviews, and business rules. These approaches typically rely on well-formed sentences and domain-specific vocabulary. For instance, SBVR-based methods depend on clearly defined fact types and logical connectors [25], while user story approaches require structured input and semantic templates [8]. In contrast, real-world user feedback is often unstructured, informal, and ambiguous, posing challenges for traditional models [8], [33]. These methods tend to perform poorly when faced with noisy input.

The proposed method addresses this gap by integrating large language models, fuzzy string matching, and syntactic patterning to extract BPMN elements from free-form text. Unlike prior approaches that depend on structured templates or curated rule sets, our method offers a flexible and scalable mechanism for modeling user-driven business processes in practical settings. This study addresses the above gap by introducing a method that integrates NLP techniques, LLM, and fuzzy string matching to translate unstructured user feedback into BPMN diagrams. The main contributions are as follows: (1) An end-to-end translation pipeline capable of processing diverse sentence structures without prior manual structuring, (2) A syntactic pattern–based preprocessing stage to improve interpretation of informal text before BPMN element extraction, (3) A comprehensive evaluation on 125 annotated feedback instances covering four sentence complexity levels (simple, compound, complex, compound-complex), and (4) The release of an annotated dataset to support reproducibility and future research in automated process modeling.

Table 1 summarizes representative studies on process model generation. Earlier works (e.g., [24]) relied on structured inputs, while later approaches [8],[33] addressed semi-structured requirements using rule-based or NLP methods. More recent efforts (e.g., [17]) employed LLMs but still depended on curated or domain-specific input formats. In contrast, the proposed FB2BPMN method uniquely handles unstructured feedback through an LLM-guided transformation with fuzzy string matching, thereby improving robustness and adaptability in processing noisy textual data. This advancement clearly fills the research gap left by previous approaches and establishes a foundation for scalable, data-driven process modeling.

The remainder of this paper is structured as follows: Section II discusses the methods, Section III presents the results and discussion, and Section IV provides the conclusion.

2 Feedback-to-BPMN method

This section describes the FB2BPMN (Feedback-to-BPMN) method developed in this study. Building on the identified research gap, the FB2BPMN framework is as an end-to-end pipeline that transforms unstructured user feedback into executable BPMN models. The approach integrates natural language processing (NLP), large language models (LLMs), and fuzzy string matching to ensure semantic alignment between informal text and formal process representations.

To guide the methodological design and evaluation, We formulated two research questions in line with the study objectives. These questions evaluate the effectiveness of the FB2BPMN pipeline in processing informal user feedback. In addition, they explore the role of fuzzy string matching in improving the consistency and semantic alignment of the generated BPMN models:

RQ1: How effectively can the FB2BPMN pipeline generate BPMN representations from informal and unstructured user feedback?

RQ2: To what extent does the integration of fuzzy string matching enhance the semantic consistency and naming alignment of BPMN elements compared to using an LLM alone?

To address these questions, the subsequent subsections describe the proposed FB2BPMN translation pipeline, which consists of preprocessing, pattern-based sentence structuring, BPMN element extraction, and fuzzy-based semantic alignment.

2.1. Architecture

FB2BPMN is designed as an end-to-end method designed to accommodate informal and syntactically diverse inputs. The method systematically generates

Table 1: Comparison of representative studies on process model generation.

Study	Dataset	Method / Approach	Performance	Limitation
Tangkawarow et al. (2022)	Informal documents	ID2SBVR: Rule-based NLP for SBVR vocabulary extraction	Qualitative validation only (syntactic and vocabulary correctness)	Limited to semi-structured inputs; requires formal grammar
Nasiri et al. (2023)	User stories	Template-driven NLP	Precision ≈ 0.94	Not applicable to free-form feedback; limited semantic flexibility
Köpke & Safan (2024)	Chat-based process descriptions	LLM-based dialogue transformation	Not evaluated quantitatively	Relies on structured prompts; conversational scope
Sholiq et al. (2022)	Textual requirements	Pattern-based / rule-based NLP	Avg. F1 ≈ 0.87	Limited adaptability to informal sentences
Arshad et al. (2019)	Controlled Natural Language (CNL) conforming to SBVR specifications	SBVR2XML: CNL→SBVR→XML transformation	Precision ≈ 0.949; Recall ≈ 0.925 (tool eval.)	Assumes controlled NL/SBVR; targets XML generation (not BPMN); does not address noisy, unstructured feedback
This study	User feedback (unstructured text)	FB2BPMN: NLP + LLM + fuzzy matching	Avg. F1 ≈ 0.91; Precision ≈ 0.92	Evaluation based on a representative 125-entry subset; future work includes broader domain and dataset expansion.

structured BPMN elements without requiring manual restructuring, thereby enabling organizations to leverage user feedback more effectively, as a foundation for business process improvement.

Figure 1 illustrates the role of the proposed FB2BPMN method within the broader cycle of business process improvement based on user feedback. FB2BPMN functions as a critical stage that translates informal textual input into structured process models, enabling systematic business process refinement. By positioning FB2BPMN in this context, the figure highlights its contribution as an integral component of the end-to-end improvement pipeline rather than an isolated technique. To the best of our knowledge, no prior study has directly addressed the improvement of business process models in such an integrated manner [40].

Building on this foundation, Figure 2 presents the internal workflow of FB2BPMN in greater detail. Specifically, the method operates through four sequential stages: (1) converting user feedback into structured text, (2) extracting relevant fact types, (3) mapping these fact types to process constructs, and (4) generating BPMN elements.

Each stage incrementally transforms unstructured input into a formal process representation, ensuring syntactic consistency and semantic interpretability. This stepwise procedure provides a clear methodological pathway for translating diverse feedback into BPMN models that can directly support business process redesign.

The method relies on LLM to analyze large volumes of previously unseen text and capture contextual nuances in user feedback [41]. By extracting semantic elements such as actions, roles, and objects, the LLM enhances the interpretation of unstructured input and addresses challenges of linguistic variability, implicit references and fragmented expressions. Rather than a standalone generator, the LLM functions as a semantic filter that enriches and disambiguates user inputs before subsequent processing.

Complementing this, NLP techniques and fuzzy string matching are employed to identify process elements through fact type extraction and to map semantically related terms to BPMN constructs [31], [42]. This combination effectively addresses both lexical variation and semantic similarity, thereby supporting the reliable generation of BPMN elements from highly diverse feedback data

2.2. Step 1: Conversion of user feedback into structured text

Algorithm 1 illustrates how a large language model (LLM), such as OpenAIGenerator, is employed with the PromptBuilder class from the Haystack library to transform raw text into sentences that follow specific syntactic patterns. Each sentence is first split into individual units (line 1) and checked for coordinating conjunctions and clauses (line 2), which determines its classification as compound (line 3), compound complex (lines 4–6), complex (line 8), or simple (line 10). These types are then mapped to corresponding prompt templates (line 12) that guide the model in generating structured representations (lines 13–14). Valid outputs are stored (lines 15–16), while invalid ones are logged (lines 17–18)

before the final set of structured sentences is returned (line 21).

```
Algorithm 1: Conversion of User Feedback into Structured Text
Data: T = user feedback text
Result: ST = structured text {Actor, Action, Object, Condition}
1.  for each sentence ∈ splitSentences(T) do
2.      if contains(sentence,["and","but","or"]) and containsClause(sentence) then
3.          type ← "compound"
4.      else if contains(sentence,["if","although","because","when","while"]) then
5.          if containsConjunction(sentence, ["and", "but", "or"]) then
6.              type ← "compound-complex"
7.          else
8.              type ← "complex"
9.          end
10.     else
11.         type ← "simple"
12.     end
13.     prompt ← TEMPLATE[type].Replace({query ← sentence})
14.     structuredSentence ← LLM.generate(prompt)
15.     if isValid(structuredSentence) then
16.         ST.add(structuredSentence)
17.     else
18.         logError("Invalid structured sentence for input: " + sentence)
19.     end
20. end
  // Each sentence tᵢ ∈ T is transformed into a structured sentence sᵢ using a
  // type-specific template.
  // ST = f_LLM(T) = { sᵢ ∈ TEMPLATE(tᵢ) | tᵢ ∈ T, type(tᵢ) ∈ {simple, compound,
  // complex, compound complex} }
21. return ST
```

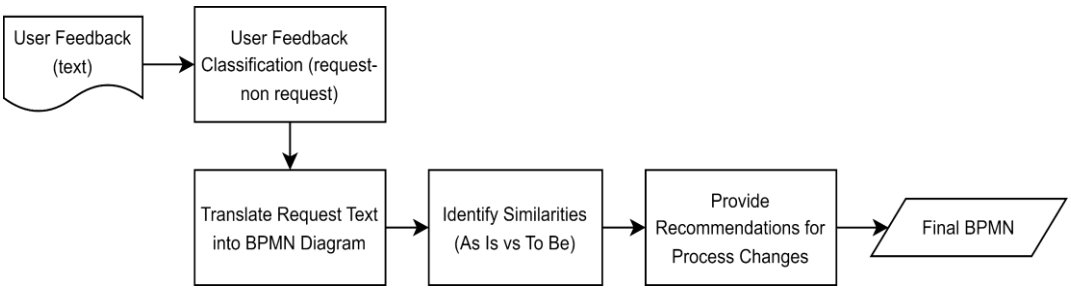


Figure 1: Overall architecture of the integrated process improvement approach.

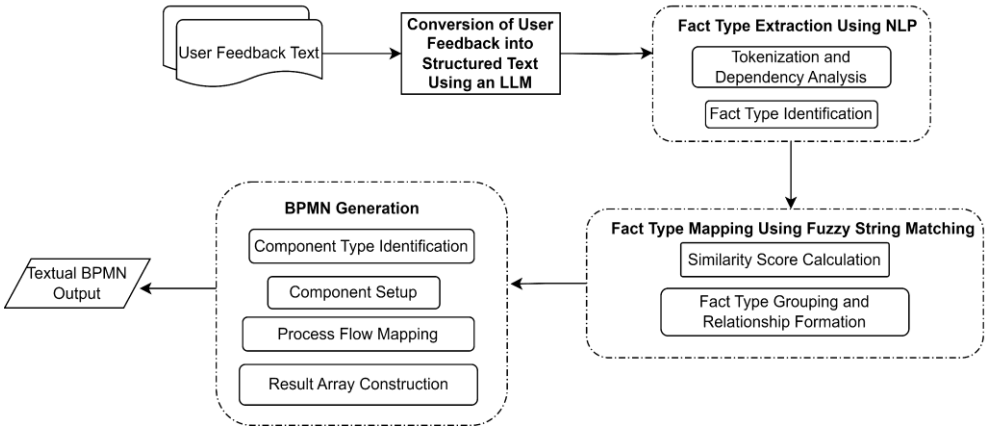


Figure 2: The FB2BPMN framework

Within the FB2BPMN method, this structured text is subsequently used for fact type extraction, mapping, and constructing BPMN elements. Figure 2 presents the architecture that supports this text to BPMN translation. Identifying such structures is essential for downstream processing, as compound and complex sentences often introduce greater syntactic ambiguity [43]. Sentence type classification reduces ambiguity and improves fact type extraction accuracy, particularly for compound and complex sentences.

Each sentence type is mapped to a specific prompt template that guides the LLM in generating a well-structured input version. For example, a simple sentence such as "*Allow tagging of articles*" might be processed

using a template: Identify the actor, action, and object in the following instruction: {sentence}, producing the structured form [System – allows – tagging of articles]. Similarly, a complex sentence such as "*If the user is logged in, allow tagging of articles*" would use a conditional template to produce separate condition and activity structures. These structured outputs improve downstream fact extraction accuracy by reducing syntactic ambiguity.

To improve clarity, the templates omit unnecessary introductory phrases. The algorithm classifies each sentence, selects the appropriate template, and submits it to the LLM. If the output meets structural criteria, it is retained; otherwise, it is flagged for review. For example, the model produces coherent and meaningful sentences reflecting the original intent from an unstructured input containing fragmented expressions and informal phrasing. While minor stylistic inconsistencies may occur, the transformation effectively enhances the readability and supports further processing in the BPMN generation pipeline.

The LLM was guided using a structured prompting template designed to transform informal feedback into syntactically consistent sentences. The prompt specified three output patterns: activity, conditional, and parallel sentences, each following predefined templates (e.g., "<subject> <verb> <object>" or "If <subject> <condition>, <activity>"). The model was instructed to rewrite user feedback from a system-centric perspective, ensuring the use of formal English, consistent grammatical subjects, and clear syntactic relations. This guided transformation produced normalized text that could be systematically parsed for fact type extraction in subsequent steps.

To illustrate, consider the input sentence "*Authors cannot track the review progress easily after submission*." The corresponding structured output generated by the LLM is {Actor: *authors*, Action: *track*, Object: *review progress*, Condition: *after submission*}.

This step classifies sentences according to their grammatical type and applies a corresponding type-specific prompt template to guide the LLM transformation. The resulting output is a normalized structured representation that preserves the original meaning while improving readability and syntactic consistency for subsequent BPMN element extraction in Algorithm 2.

The complete LLM prompt template used in this step is provided in the supplementary material available on Figshare at https://doi.org/10.6084/m9.figshare.29023172.

2.3. Step 2: fact type extraction

This step analyzes the structured text and categorizes sentences into two primary fact types: activities and conditions. These categories form the basis for mapping textual information to BPMN components, such as tasks and gateways. The process comprises two subprocesses: (1) tokenization and dependency analysis, and (2) fact type extraction.

The first subprocess, tokenization and dependency analysis, involves breaking each sentence into tokens (words) and labeling their grammatical roles, such as subject, predicate, and object. This analysis uses the spaCy Python library, which provides a detailed grammatical structure for each sentence. For example, the sentence "*The system processes user data*" is parsed into tokens: "The," "system," "processes," "user," and "data", with dependencies identified as "system: Subject," "processes: Predicate," and "data: Object." This step ensures that the system can accurately capture syntactic relationships within the text.

Sentences that lack a complete subject predicate object structure are flagged for manual review or passed through a secondary LLM based reconstruction. For instance, the incomplete feedback "*Need option for article tagging*" omits an explicit subject; the secondary pass infers "*System*" as the default actor before proceeding to extraction. This dual pass mechanism minimizes data loss and maintains the integrity of the fact type dataset.

This procedure is summarized in Algorithm 2, which integrates tokenization, dependency analysis, and fact type extraction. The algorithm processes each structured sentence to extract the subject, action, and object, forming the grammatical basis for classification as either activity or condition fact types.

To illustrate, consider the structured input {Actor: *authors*, Action: *track*, Object: *review progress*, Condition: after submission}. From this representation, the algorithm extracts two fact types: activity (*authors, track, review progress*) and condition (*after submission*).

This step converts structured sentences into semantic triples that capture either actions or conditions, providing the input for Algorithm 3.

2.4. Step 3: fact type mapping

This step applies fuzzy string matching, using the Levenshtein distance metric, to align text segments that are similar but not identical. The process consists of two subprocesses: (1) identifying similarities between fact types, and (2) generating sequence flows based on the relationships inferred from the similarity scores, as illustrated in Algorithm 3.

Fuzzy string matching is applied to align semantically similar fact types that differ in wording. This study uses the Levenshtein distance metric, with a similarity threshold of 0.85. This value was selected empirically to balance sensitivity (capturing legitimate variations) and specificity (avoiding incorrect merges). Although alternative measures such as Jaro–Winkler and cosine similarity were considered, Levenshtein distance was chosen for its robustness to short text variations and token-level edits, common in user feedback.

The Levenshtein distance measures the minimum number of operations required to transform one string into another, namely deletions, insertions, or substitutions. It is computed by initializing a matrix to represent the prefix lengths of both strings and filling it with the minimal number of operations needed at each position. The

```
Algorithm 2: Fact Type Extraction
Data: ST = structured text produced by Algorithm 1
Result: FT = Extracted fact types
for each sentence ∈ structuredText do
  //Tokenization and dependency analysis
  tokens ← tokenize(sentence)
  dependencyLabels ← parseDependency(tokens)
 // Fact type extraction
if containsKeyword(tokens, "if")then
  condition ← extractCondition(sentence)
  factType ← createFact("condition", condition)
 else
 //Identify core elements
   subject ← findTokenByLabel(dependencyLabels, "nsubj")
   verb ← findTokenByLabel(dependencyLabels, "ROOT")
  object ← findTokenByLabel(dependencyLabels, "dobj")
 if subject ≠ null and verb ≠ null and object ≠ null then
   activity ← createActivity(subject, verb, object)
   factType ← createFact("activity", activity)
 else
   logWarning("Incomplete sentence structure: " + sentence)
   continue
  end
 end
   factTypes.append(factType)
end
  // Each structured sentence sᵢ ∈ ST is mapped to a fact type fᵢ
  // FT = f_extract(ST) = { fᵢ ∈ MAPPED(sᵢ) | sᵢ ∈ ST }
return factTypes
```

resulting distance is converted into a similarity score (simScore) using the formula presented in Eq. (1)

$$simScore(x_i, x_j) = \left(1 - \frac{distance(x_i, x_j)}{\max\text{len}(x_i, x_j)}\right) \times 100 \qquad (1)$$

As shown in Eq. (1), the distance (x_i, x_j) measures dissimilarity or difference between x_i and x_j, and maxlen (x_i, x_j) is the length of the longer sequence between x_i and x_j. For example, given the sequences x_i = “user” and x_j = ”use”, we can calculate that the distance (x_i, x_j) is the number of edits required to align the sequences. In this case, deleting 'r' from "user" results in distance (x_i, x_j) = 1, while maxlen (x_i, x_j) represents the length of the longer sequence, in this case "user," yielding maxlen = 4. Therefore, the similarity score between x_i and xj is computed based on these values.

Similarity scores are used to identify relationships among fact types. Fact types exhibiting high similarity are consolidated into a single BPMN element, whereas those with low similarity are treated as distinct entities. The BPMN model encapsulates clusters of similar fact types and represents their interrelations through sequence flows. Contextually related but distinct elements are connected accordingly, while highly similar ones preserve their original source and target associations. This step employs a fuzzy string matching algorithm to analyze and map semantic relationships across fact types.

To illustrate, consider the following input fact types derived from Algorithm 2: activity (authors, track, review progress) and condition (after, submission). The temporal cue after submission establishes a dependency relation between the two facts, indicating that the activity occurs subsequent to the condition. Consequently, the algorithm produces a relation in which the completion of submission precedes the activity of tracking the review progress.

This step identifies semantic and temporal relationships among extracted fact types, which serve as the foundation for BPMN element generation in Algorithm 4.

2.5. 4: BPMN generation

In this step, two subprocesses, namely generating BPMN elements and mapping relationships, construct BPMN models from fact types and their semantic links. Fact types are first transformed into BPMN tasks or gateways, depending on whether they represent activities or conditions. The relationships between these components are then mapped using sequence or conditional flows, ensuring logical coherence in the resulting model. A textual representation of the BPMN

The algorithm in Step 4 consists of two subprocesses: generating BPMN elements and mapping relationships. Activity fact types are transformed into tasks, and condition fact types into gateways, while relationships between them are mapped using sequence or conditional flows. These subprocesses result in a textual BPMN representation supporting visualization and implementation, as illustrated in Algorithm 4. The resulting BPMN XML adheres to the BPMN 2.0 specification, enabling direct import into standard modeling tools such as bpmn.io or Camunda Modeler.

To illustrate, consider the following example derived from the previous step. The input fact types are (*authors, track, review progress*) and (*after submission*), with the relation (*submission - track review progress*). From these inputs, the algorithm generates BPMN elements in both XML and JSON formats, as shown below.

Generated BPMN elements (XML):

<bpmn:task id="T1" name="Submit Manuscript"/>
<bpmn:exclusiveGateway id="G1" name="After Submission"/>

```
Algorithm 3: Fact Type Mapping
Data: FT = extracted fact types
Result: R = relationships between fact
  //Identify similarities between fact types
1.for each factType1 ∈ factTypes do
2. for each factType2 ∈ factTypes do
3.  if factType1 ≠ factType2 then
4.   score ← fuzzy_match(factType1.name,factType2.name)
5.    if score ≥ threshold then
6.     relation ← createRelation(factType1,factType2)
7.    relations.append(relation)
8. end
9.end
10.for each relation ∈ relations do
  // Generate sequence flows from relationships
11. source ← relation.source
12. target ← relation.target
13. sequenceFlow ← createSequenceFlow source.id, target.id)
14.bpmnSequenceFlows.append(sequenceFlow)
15.end
  // Formal representation:
  // R = f_map(FT) = { (f_i, f_j) | sim(f_i, f_j) ≥ θ, f_i ≠ f_j }
  // Where sim is a fuzzy similarity function and θ is the threshold
16.return bpmnSequenceFlows
```

```
Algorithm 4: BPMN Generation
input: FT = Extracted fact type, R = Relationship between facts types
Result: BPMN textual, BPMN XML
  //Genareating BPMN elements
1.Initialize BPMN_textual ← []
2.Create BPMN_XML root element (e.g.,<bpmn:definitions>)
3.Add <bpmn:process> element to BPMN_XML
4.for each factType ∈ FT do
5. if factType.type == "activity" then
6.  task ← createTask(factType.subject, factType.action, factType.object)
7.  BPMN_textual.add(task)
8.  Add <bpmn:task> to BPMN_XML with task attributes
9. else if factType.type == "condition" then
10.  gateway ← createGateway factType.condition)
11.  BPMN_textual.add(gateway)
12.  Add <bpmn:exclusiveGateway> to BPMN_XML with gateway attributes
13.  end
14.end
   //Mapping relationships
15.for each relation ∈ R do
16.  if relation.type == "dependency" then
17.    flow ← createSequenceFlow (relation.source, relation.target)
18.   BPMN_textual.add(flow)
19.   Add <bpmn:sequenceFlow> to BPMN_XML connecting source and target
20.  else if relation.type == "condition" then
21.   flow ← createConditionalFlow (relation.source, relation.target)
22.   BPMN_textual.add(flow)
23.   Add <bpmn:sequenceFlow> with condition to BPMN_XML
24.  end
25. end
    // BPMN = f_generate(FT, R) = {elements ∪ flows|elements from FT, flows from R}
26. return BPMN
```

<bpmn:task id="T2" name="Track Review Progress"/>
<bpmn:sequenceFlow sourceRef="T1" targetRef="T2"/>

Generated BPMN elements (JSON):

{"bpmn_shapes": [{"id": 0, "type": "task", "name": "Submit Manuscript"}, {"id": 1, "type": "gateway", "name": "After Submission"}, {"id": 2, "type": "task", "name": "Track Review Progress"}], "sequence_flows": [{"source": 0, "target": 1}, {"source": 1, "target": 2}]}

This step converts semantic fact types and their identified relationships into BPMN-compliant elements in both textual and XML representations, completing the FB2BPMN transformation pipeline.

3 Results and discussion

This section presents the results of translating user feedback into BPMN models (BPMN1) using the proposed method and compares them with expert-authored reference models (BPMN2). The evaluation covers dataset composition, experimental procedure, and performance analysis based on standard metrics, followed by an interpretation of the findings and their implications.

3.1. Experiments

A total of 1,027 user feedback entries were initially collected from the Public Knowledge Project's Open Journal Systems (PKP OJS) platform. These entries comprised diverse forms of user-generated text, including feature suggestions, bug reports, and general comments. Following expert annotation, entries that did not contain explicit requests for new features or improvements to

existing functionalities were excluded, resulting in a refined dataset of 656 relevant feedback instances.

The dataset exhibits substantial structural variability, ranging from 30 to 50 words and from 2–5 sentences per entry. Typical noise patterns include @-mentions, URLs, issue or pull request references, markup tags (e.g., <h2>), and interface-specific tokens such as menu labels or configuration paths. These heterogeneous characteristics motivated the design of the sentence structuring process (Algorithm 1) and subsequent fact extraction and alignment steps designed to handle noisy and irregular input.

From this filtered dataset, 125 entries were selected using stratified random sampling to ensure proportional representation of major sentence types. The sample size was determined by practical constraints, as manual evaluation required significant domain expertise and effort. Nevertheless, the selected subset provides sufficient linguistic diversity to support a reliable assessment of the proposed FB2BPMN method.

All 125 entries were successfully processed, demonstrating the system's capability to interpret unstructured and variable user input. The dataset includes four sentence types: simple, compound, complex, and compound–complex, providing a representative foundation for evaluating the method's effectiveness in translating natural language into structured BPMN models. Prior studies have shown that sentence complexity can significantly affect performance in downstream transformation tasks [43].

The final sample consists of 28 complexes, 33 compound, 14 compound-complex, and 50 simple sentences, corresponding to 22.4%, 26.4%, 11.2%, and 40% of the dataset. This distribution ensures that each sentence type is adequately represented, allowing for a nuanced evaluation of the method's robustness across varying levels of syntactic complexity. The supplementary dataset, including all 125 annotated instances and evaluation outputs, is available in the Figshare repository (DOI: 10.6084/m9.figshare.29023172).

Each feedback instance was translated into structured text and then converted into BPMN models (BPMN1), which were evaluated against expert-generated reference models (BPMN2). The comparison was conducted based on three main evaluation criteria: (1) element identification, which assesses the completeness of BPMN2 elements relative to BPMN1; (2) element naming, which evaluates terminological consistency; and (3) process flow, which examines the alignment of activity sequences. Performance was measured using standard metrics, including accuracy, precision, recall, and F1 score, to assess the model's effectiveness in producing BPMN representations that accurately reflect user feedback.

Each feedback case was processed individually, and in certain instances, multiple translation iterations were applied to improve contextual fidelity. For example, the feedback "currently dependent files can only be added/administered on the galleys grid. Allow them elsewhere" as refined through several iterations to capture the intended process semantics more accurately. The expanded evaluation using 125 datasets enabled a broader and more representative analysis of system performance across varying levels of feedback complexity.

Table 2: The distribution of sentence types.

Sentence Type	Dataset ID	Entries
Complex	1, 9, 21, 43, 58, 76, 81, 85, 95, 97, 98, 99, 101, 102, 104, 105, 106, 107, 108, 110, 111, 112, 114, 115, 116, 117, 118, 124	28
Compound	10, 13, 15, 24, 32, 35, 36, 38, 39, 40, 41, 42, 44, 46, 47, 49, 51, 55, 57, 61, 63, 68, 71, 72, 75, 77, 83, 84, 86, 89, 91, 92, 93	33
Compound-Complex	45, 62, 80, 90, 94, 96, 100, 103, 109, 119, 120, 121, 123, 125	14
Simple	2, 3, 4, 5, 6, 7, 8, 11, 12, 14, 16, 17, 18, 19, 20, 22, 23, 25, 26, 27, 28, 29, 30, 31, 33, 34, 37, 48, 50, 52, 53, 54, 56, 59, 60, 64, 65, 66, 67, 69, 70, 73, 74, 78, 79, 82, 87, 88, 113, 122	50

Table 2 summarizes the distribution of sentence types within the dataset, providing a representative foundation for evaluating the performance of the proposed method in the subsequent section.

3.2. Evaluation

The evaluation aims to measure the translation model's performance using standard metrics, including accuracy, precision, recall, and F1 score [8], [44], [45]. It compares the reference textual BPMN with the generated textual BPMN from the translation process using the same dataset.

The dataset was derived from a platform that developed an Open Journal System (OJS), containing user feedback related to business process improvement requests. Reference textual BPMN models (BPMN2) were manually created by experts, specifically online journal administrators at a university. The generated models (BPMN1) were produced by the proposed translation method. Structured feedback text was provided to the experts, along with a BPMN writing template aligned with the array-based data structure of the generated BPMN. This template ensured a uniform format for comparison. The evaluation focused on three aspects: element identification, element naming, and process flow:

(1) Element Identification (tasks, gateways): Does BPMN2 include all the elements that appear in BPMN1?
(2) Element Naming: Are there discrepancies in the naming of elements? This assessment considers exact matches and semantic similarity, measured against a predefined threshold. For example, "if it meets the standards" and "if it meets the basic standards" would

be considered equivalent if their meanings are aligned, despite minor differences in phrasing.

(3) Process Flow: Does the sequence of activities in BPMN2 follow the same logical order as in BPMN1?

The following definitions are applied to calculate the evaluation metrics, namely accuracy (Acc), precision (Prec), recall (Rec), and F1 score (F1). True Positive (TP) refers to elements that correctly appear and match in BPMN2 as expected based on BPMN1, including both the name and the sequence. False Positive (FP) refers to additional or mismatched elements in BPMN2 that do not exist in BPMN1 or are incorrect. False Negative (FN) refers to elements that should be present according to BPMN1 but are missing or incorrect in BPMN2. For example, in one feedback case, the generated BPMN1 contained the task "Allow tagging of articles" and the gateway "If the user is logged in", both of which also appeared in BPMN2. These were counted as true positives. However, BPMN2 included an additional task "Display confirmation message" not present in BPMN1 (false negative), while BPMN1 included "Check user role" which was absent in BPMN2 (false positive).

Table 4 presents the average performance of the FB2BPMN system across four sentence types: simple, compound, complex, and compound complex, evaluated using accuracy, precision, recall, and F1 score. The assessment covers three fundamental BPMN modeling tasks: element identification (A), element naming (B), and process flow construction (C).

Sentences with greater syntactic complexity, particularly complex and compound complex types, consistently yield higher scores in element identification and process flow tasks. This trend is likely driven by richer syntactic cues and more explicit relational structures. In contrast, simple sentences have the lowest scores, especially in the naming task, due to limited contextual and lexical information.

Table 3: Summary statistics of evaluation metrics

Aspect	Accuracy	Precision	Recall	F1-score
Mean:				
Identification	0.87	0.97	0.88	0.91
Naming	0.77	0.91	0.77	0.83
Process Flow	0.85	0.96	0.86	0.90
Std. Deviation:				
Identification	0.15	0.08	0.15	0.11
Naming	0.17	0.15	0.16	0.14
Process Flow	0.19	0.10	0.19	0.13

Among the three tasks, element naming (B) demonstrates the greatest variability across metrics. Precision decreases for complex and compound complex sentences, indicating challenges in assigning consistent labels when multiple clauses or entities are present. Meanwhile, process flow performance (C) improves with sentence complexity, as reflected in higher F1 scores, suggesting that complex syntactic structures support more accurate reconstruction of procedural logic.

These results have direct practical implications. As presented in Table 3, high precision in element identification (97%) suggests the method is reliable when accuracy is critical, such as in safety-critical domains or compliance auditing. However, lower recall in element naming (77%) indicates that the method may omit some valid elements or produce inconsistent labels, which could require post-processing or human validation in applications demanding high completeness. Overall, the results indicate that increased sentence complexity enhances the system's ability to extract structured representations, leading to more accurate and coherent BPMN model generation.

The results presented in Table 3 indicate that the model performs well in satisfying syntactic expectations across key evaluation dimensions. Metrics were obtained through a fully manual evaluation, in which the BPMN elements generated by the model were compared against expert-labeled ground truth. The assessment focused on semantic adequacy, label appropriateness, and contextual relevance, ensuring that the evaluation captured structural accuracy, semantic fidelity, and domain-specific validity. Preliminary findings suggest that the model's high precision in element identification and flow construction aligns with human judgment. However, variations remain in the naming aspect, where semantic interpretation often diverges due to paraphrasing or lexical ambiguity in user inputs. These results underscore the need for further qualitative assessment to better understand the trade-off between syntactic accuracy and semantic fidelity in real-world scenarios.

While Table 3 summarizes the overall performance across identification, naming, and process flow aspects, further analysis was conducted to examine how sentence structure influences these outcomes. The goal of this complementary evaluation was to assess whether syntactic complexity affects the model's ability to accurately extract BPMN elements from unstructured feedback.

As shown in Table 4, the FB2BPMN pipeline maintained consistent performance across all sentence types, with only a slight decrease observed for compound-complex sentences due to increased syntactic ambiguity.

Table 4: Performance by sentence type

Sentence Type	Precision	Recall	F1-Score
Simple	0.95	0.90	0.92
Compound	0.94	0.89	0.91
Complex	0.93	0.88	0.90
Compound-Complex	0.91	0.86	0.88
Average	0.93	0.88	0.90

To illustrate specific challenges, Dataset 14 achieved perfect element identification (1.00), yet naming recall declined to 0.60 due to lexical variation. For example, BPMN1 generated "*Export all button will be applied*" and "*Select all option would be preferable on a single grid page*," whereas BPMN2 produced "*Export all button will be added*" and "*Select all option is preferred for single grid pages.*" Although semantically equivalent, these differences reduced the naming score despite minimal

impact on process logic. Dataset 35 demonstrated structural compression, where BPMN1 specified four distinct activities ("*The system provides a content editor for HTML versions," "The system saves HTML versions to the database or as flat files," "Some journals request this functionality*," *and "The system acknowledges the difficulties that this functionality presents*"). At the same time, BPMN2 abstracted them into two generalized tasks ("*Provide and save HTML content editor output" and "Handle journal requests and acknowledge editor difficulties*"). This abstraction reduced recall for both identification and naming (approximately 0.50–0.60) and lowered flow accuracy to 0.33, preserving overall logic but limiting granularity and traceability.

In contrast, Datasets 64 and 65 achieved full alignment across all metrics, confirming robustness when the input was unambiguous. For instance, in Dataset 64, BPMN1 produced tasks such as "Authors upload revised figures separately" and "Allow partial update of figures only," with a gateway "Should this happen without changing full submission?"; BPMN2 generated semantically equivalent elements "Let authors upload revised figures separately," "Simplify figure-only updates in workflows," and a gateway "Does this avoid full resubmission?" Likewise, in Dataset 65, BPMN1 specified "Import metadata from external databases" and "Reduce errors in manual metadata entry," with a gateway "Can this reduce manual entry errors?"; BPMN2 produced corresponding elements "Enable metadata import from external sources," "Reduce manual input errors," and a gateway "Does this minimize entry mistakes?" All elements were consistently identified, named, and sequenced, yielding complete alignment between model outputs and expert references. These examples highlight two recurring sources of error: omission or merging of elements, which primarily affects recall and flow reconstruction, and semantic compression, where multiple activities are combined into broader elements, thereby reducing naming precision and structural detail. While the generated BPMN models preserve high-level process logic, further improvements are required to support use cases demanding high semantic fidelity and fine-grained process validation. Complete results for all 125 cases are provided in the Figshare repository.

An aggregated evaluation complemented the qualitative analysis and examined overall consistency. Table 5 presents a comparative summary of BPMN evaluation results using standard classification metrics (Precision, Recall, F1, Accuracy). The rule-based baseline method [32] was analyzed separately, as it employed MMRE-based error evaluation rather than element-wise classification metrics. The reported values represent the mean and standard deviation of accuracy, precision, recall, and F1-score across all datasets, confirming the model's robustness under varying syntactic and semantic complexity levels. The complete dataset-level evaluation table, including detailed metric values for all 125 instances, is provided in the Figshare repository for transparency and reproducibility. A comparative analysis with existing BPMN generation approaches is presented to contextualize these findings. Although the overall accuracy achieved by the proposed method (87%) is slightly lower than that reported by [8] (94%), it demonstrates higher precision, reaching 97% in element identification. It reflects a conservative extraction strategy that minimizes false positives, making the approach suitable for scenarios that require high-confidence mapping from natural language feedback. Unlike Nasiri's method [8], which assumes structured and well-formed input, FB2BPMN prioritizes robustness under noisy and unstructured conditions. This design choice emphasizes high precision while trading off some recall, ensuring reliability in real-world scenarios where feedback is highly variable.

To assess computational efficiency, we evaluated the average runtime of each component using Google Colab Pro with an NVIDIA T4 GPU and 32 GB RAM. The sentence restructuring phase completed within a few seconds per input on average, while fact extraction and mapping were executed even faster. BPMN diagram generation was near-instantaneous due to its template-driven logic. These observations indicate that the method provides consistent and responsive performance, ensuring practical scalability for processing user feedback in real-world settings.

To further assess the effectiveness of the proposed method, we provide a qualitative comparison with existing BPMN generation approaches that rely on structured inputs. For example, [8] proposed a user story-based approach that utilizes predefined semantic patterns. Danenas et al. [25] extracted SBVR-based elements from UML use case diagrams using natural language processing techniques. These methods typically require domain-specific templates and are limited in handling informal or ambiguous user expressions.

In contrast, our method demonstrates greater robustness in processing unstructured user feedback, which frequently contains irregular syntax, vague references, and multiple intents. While structured input methods often fail to generate valid BPMN components without extensive preprocessing, our approach achieves up to 95 percent accuracy in element identification, even when applied to free-form text.

Although a direct quantitative comparison was not feasible due to differing input assumptions and data representations, the proposed method demonstrates clear advantages in handling informal and unconstrained input sources, broadening the practical applicability of automated business process modeling. The findings suggest that the method is particularly suited for domains where feedback is abundant but formal documentation is scarce, such as public service systems, e-commerce

Table 5: Comparative summary of BPMN evaluation methods.

Aspect	This Study	Nasiri et al., [8]
Evaluation Method	Precision, recall, F1 score, and accuracy	Precision, recall, and accuracy based on true/false positive and negative counts
Evaluation Focus	Element detection, naming consistency, and process flow correctness	Activity correctness, control flow validation, and transition accuracy
Key Results	Precision = 97%, Recall = 88%, Accuracy = 87%	Precision = 94%, Recall = 97%, Accuracy = 94%
Strengths	Reliable element identification under informal and variable input structures	High semantic accuracy in modeling control flow and behavioral transitions

platforms, and citizen reporting tools. In such contexts, generating high-precision process models from informal text can accelerate process redesign, facilitate bottleneck identification, and reduce reliance on domain experts for initial modeling.

The proposed method shows strong potential for application in real-world settings, particularly within public service systems, customer support platforms, and business process improvement teams. By enabling the transformation of informal user feedback into structured BPMN representations, the method allows business analysts to identify recurring issues, recommend service redesigns, and automate routine modeling activities. For example, in a government complaint handling system, free-text reports from citizens could be converted into process models to reveal bottlenecks or procedural inefficiencies. Recent studies further support the practical viability of this approach. Licardo et al. [31] demonstrated that BPMN models can be accurately extracted from informal textual descriptions using a combination of large language models and natural language processing techniques, achieving up to 96 percent accuracy. Furthermore, a systematic review by [46] highlights the increasing effectiveness of machine learning and deep learning methods, particularly LLM-based approaches, compared to traditional rule-based techniques when processing unstructured and fragmented business process descriptions. These findings affirm the applicability of our method in handling real-world input scenarios where formal documentation is often lacking. This ability to extract structured process knowledge from unstructured narratives supports evidence-based decision making and promotes continuous improvement in dynamic organizational environments.

As presented in Table 5, the proposed method achieves an overall accuracy of 87%, significantly lower than the 94% reported by [8]. However, it demonstrates higher precision, particularly in element identification (97%, see Table 3). This trade-off indicates that the method identifies fewer total elements but with greater consistency, reducing false positives. Such precision-oriented behavior is especially valuable in domains where model clutter or misinterpretation must be minimized. In contrast, Nasiri's method [8] yields higher accuracy but at the cost of lower precision, increasing the risk of redundant or inconsistent elements.

These findings also highlight several limitations that inform future research directions. First, the variability in naming accuracy indicates the need for semantic enhancement, such as contextual embeddings or domain-specific ontologies, to improve consistency. Second, while the method demonstrates high precision, the trade-off with recall suggests that hybrid modeling strategies could better balance granularity and completeness. Finally, since the evaluation was conducted on a single dataset, cross-domain validation is required to assess the generalizability and adaptability of the method across different organizational contexts.

These results answer RQ1, indicating that large language models can effectively process informal and unstructured user feedback to generate accurate BPMN representations. The FB2BPMN pipeline maintained consistent precision and recall across all four sentence types: simple, compound, complex, and compound-complex, demonstrating robustness in handling syntactic variability within unstructured text.

Regarding RQ2, the integration of fuzzy string matching notably improved naming consistency and semantic alignment in BPMN generation compared to rule-based baselines. This enhancement reduced synonym-related naming variations and increased overall correspondence between user-expressed actions and BPMN activity labels, confirming the benefit of combining fuzzy matching with LLM-based extraction.

For reference, the rule-based method by [32] provides a suitable baseline for BPMN generation from textual requirements. Their rule and pattern-based framework achieved an average F1 score of about 0.87 on structured requirement documents. In comparison, the proposed FB2BPMN method reached an average F1-score of 0.92 on unstructured user feedback, highlighting the benefit of combining large language models with fuzzy matching to handle informal, domain-specific text.

Beyond quantitative gains, FB2BPMN shows strong potential for real-world deployment across domains such as customer service, healthcare, and industrial workflows, where unstructured feedback can be transformed into actionable process models.

The proposed FB2BPMN framework can also be applied to healthcare systems, where user feedback often highlights operational or service-related issues. For instance, the feedback "*It's difficult to schedule an appointment with the same doctor again after a consultation. Please allow users to rebook easily and add a reminder for follow-up visits.*" is transformed into structured sentences: "*The system allows users to rebook appointments easily*." and "*If a consultation is completed, the system provides a reminder for follow-up visits*." From these, FB2BPMN extracts the fact types (system, allow, rebook appointment) and (consultation, completed),

forming a temporal dependency from consultation to rebook appointment. The resulting BPMN represents a patient follow-up workflow, showing that FB2BPMN effectively captures procedural logic from healthcare narratives with minimal domain adaptation.

A full expert-based evaluation has also been completed, with results openly available in the accompanying Figshare repository.

Despite these promising results, certain threats to validity remain. Although this study used 125 OJS feedback instances, the method demonstrated robust performance across diverse sentence types. Future work will extend the evaluation to larger or multilingual datasets. Conceptually, the approach exhibits robustness in processing noisy and uncertain language, analogous to adaptive fuzzy control mechanisms that stabilize nonlinear systems under uncertainty [47]. These results indicate both the practical value and theoretical soundness of the FB2BPMN framework, laying the groundwork for broader domain and cross-lingual applications.

4 Conclusions

This study presents a method that integrates Natural Language Processing (NLP), Large Language Models (LLM), and fuzzy string matching to automatically translate unstructured user feedback into Business Process Model and Notation (BPMN) diagrams. Addressing a critical gap in business process modeling, the method systematically transforms informal textual input into structured BPMN representations through a four-stage methodology: (1) text structuring, (2) fact extraction, (3) relationship mapping, and (4) BPMN generation. This design supports processing linguistically diverse inputs without requiring prior manual structuring.

Evaluation on 125 annotated feedback instances demonstrated that the proposed method performs robustly across varying levels of feedback complexity. It achieves mean accuracies of 87% for element identification, 77% for naming, and 85% for process flow mapping, with precision scores of 97%, 91%, and 96%, respectively. This high level of precision is particularly valuable in real-world applications where output reliability is critical.

Despite these strengths, challenges remain in achieving consistent element naming and handling semantically diverse expressions, particularly in feedback containing multiple intertwined intents or high lexical variability. Moderate standard deviations observed in naming and flow metrics suggest opportunities for refinement, especially in handling lexical variation and feedback containing multiple intents. Further investigation is needed to improve semantic alignment and address abstraction and compression issues in multi-topic feedback. Additionally, sentence structure was found to influence model performance. Complex and compound complex sentences yielded the highest accuracy and F1 scores across all evaluation aspects, suggesting that syntactic richness enhances interpretability and translation quality. These findings underscore the importance of linguistic features in optimizing user-to-model transformation workflows.

The method establishes a reliable and scalable foundation for translating informal user feedback into BPMN models. It contributes to the development of intelligent, user-responsive business process modeling systems. Theoretically, this study extends BPMN generation research by bridging informal human feedback with formal process modeling through LLM-based structuring. Future work will focus on semantic enhancement, hybrid modeling, and cross-domain validation to improve consistency, balance precision and recall, and ensure generalizability.

Acknowledgement

This work was supported by Institut Teknologi Sepuluh Nopember for the research activities and by Universitas Muhammadiyah Sidoarjo for the publication funding.

References

[1] K. Grolinger, M. A. M. Capretz, A. Cunha, and S. Tazi, “Integration of business process modeling and Web services: A survey,” *Serv. Oriented Comput. Appl.*, vol. 8, no. 2, pp. 105–128, 2014, https://doi.org/10.1007/s11761-013-0138-2.

[2] T. Skersys, P. Danenas, and R. Butleris, “Extracting SBVR business vocabularies and business rules from UML use case diagrams,” 2018, https://doi.org/10.1016/j.jss.2018.03.061.

[3] V. Kale, “Business Process Modeling and Notation,” *Enterp. Process Manag. Syst.*, no. January, pp. 257–275, 2019, https://doi.org/10.1201/9780429453311-14.

[4] M. Javed and Y. Lin, “iMER: Iterative process of entity relationship and business process model extraction from the requirements,” *Inf. Softw. Technol.*, vol. 135, no. January, p. 106558, 2021, https://doi.org/10.1016/j.infsof.2021.106558.

[5] U. Indahyanti, A. Djunaidy, and D. Siahaan, “Auto-Generating Business Process Model From Heterogeneous Documents : A Comprehensive Literature Survey,” in *Proc. 9th Int. Conf. Electrical Engineering, Computer Science and Informatics*, 2022, pp. 239–243. https://doi.org/https://doi.org/10.23919/EECSI56542.2022.9946460.

[6] I. Tangkawarow, R. Sarno, and D. Siahaan, “ID2SBVR: A Method for Extracting Business Vocabulary and Rules from an Informal Document,” *Big Data Cogn. Comput.*, vol. 6, no. 4, 2022, https://doi.org/10.3390/bdcc6040119.

[7] F. Gilson, M. Galster, and F. Georis, "Generating use case scenarios from user stories," in *Proceedings - 2020 IEEE/ACM International Conference on Software and System Processes, ICSSP 2020*, 2020, pp. 31–40. https://doi.org/10.1145/3379177.3388895.

[8] S. Nasiri, A. Adadi, and M. Lahmer, "Automatic generation of business process models from user stories," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 1, pp. 809–822, 2023, https://doi.org/10.11591/ijece.v13i1.pp809-822.

[9] I. Beerepoot *et al.*, "The biggest business process management problems to solve before we die," *Comput. Ind.*, vol. 146, no. December 2022, p. 103837, 2023, https://doi.org/10.1016/j.compind.2022.103837.

[10] T. Grisold *et al.*, "The Five Diamond Method for Explorative Business Process Management," *Bus. Inf. Syst. Eng.*, vol. 64, no. 2, pp. 149–166, 2022, https://doi.org/10.1007/s12599-021-00703-1.

[11] A. Mustansir, K. Shahzad, and M. K. Malik, "Towards automatic business process redesign: an NLP based approach to extract redesign suggestions," *Autom. Softw. Eng.*, vol. 29, no. 1, 2022, https://doi.org/10.1007/s10515-021-00316-8.

[12] J. Wouters, A. Menkveld, S. Brinkkemper, and F. Dalpiaz, "Crowd-based requirements elicitation via pull feedback: method and case studies," *Requir. Eng.*, vol. 27, no. 4, pp. 429–455, 2022, https://doi.org/10.1007/s00766-022-00384-6.

[13] I. Fatawi, M. Asy'ari, H. Hunaepi, T. Samsuri, and M. R. Bilad, "Empowering Language Models Through Advanced Prompt Engineering: A Comprehensive Bibliometric Review," *Indones. J. Sci. Technol.*, vol. 9, no. 2, pp. 441–462, 2024, https://doi.org/10.17509/ijost.v9i2.71481.

[14] A. Mukanova *et al.*, "LLM-Powered Natural Language Text Processing for Ontology Enrichment," *Appl. Sci.*, vol. 14, no. 13, 2024, https://doi.org/10.3390/app14135860.

[15] P. Danenas and T. Skersys, "Exploring Natural Language Processing in Model-To-Model Transformations," *IEEE Access*, vol. 10, no. November, pp. 116942–116958, 2022, https://doi.org/10.1109/ACCESS.2022.3219455.

[16] S. Prasad, H. Gupta, and A. Ghosh, "Leveraging the Potential of Large Language Models," *Inform.*, vol. 48, no. 8, pp. 1–16, 2024, https://doi.org/10.31449/inf.v48i8.5635.

[17] R. Sonbol, G. Rebdawi, and N. Ghneim, "A Machine Translation Like Approach to Generate Business Process Model from Textual Description," *SN Comput. Sci.*, vol. 4, no. 3, pp. 1–16, 2023, https://doi.org/10.1007/s42979-023-01742-z.

[18] J. Köpke and A. Safan, "Introducing the BPMN-Chatbot for Efficient LLM-Based Process Modeling," 2024, https://ceur-ws.org/Vol-3758/paper-15.pdf.

[19] T. Fehrer, D. A. Fischer, S. J. J. Leemans, M. Röglinger, and M. T. Wynn, "An assisted approach to business process redesign," *Decis. Support Syst.*, vol. 156, no. July 2021, 2022, https://doi.org/10.1016/j.dss.2022.113749.

[20] P. Delias and G. T. Nguyen, "Prototyping a business process improvement plan. An evidence-based approach," *Inf. Syst.*, vol. 101, p. 101812, 2021, https://doi.org/10.1016/j.is.2021.101812.

[21] F. Johannsen and H. G. Fill, "Meta Modeling for Business Process Improvement," 2017, doi: 10.1007/s12599-017-0477-1.

[22] Q. Zeng, J. Liu, C. Zhou, C. Liu, and H. Duan, "A Novel Approach for Business Process Similarity Measure Based on Role Relation Network Mining," 2020, https://doi.org/10.1109/ACCESS.2020.2983114.

[23] S. Ayari, Y. B. Hlaoui, and L. Ben Ayed, "A specific language for developing business process by refinement based on BPMN 2.0," *Proc. 16th Int. Conf. Softw. Technol. ICSOFT 2021*, no. July, pp. 489–496, 2021, https://doi.org/10.5220/0010561404890496.

[24] S. Arshad, I. S. Bajwa, and R. Kazmi, "Generating SBVR-XML Representation of a Controlled Natural Language," in *Communications in Computer and Information Science*, 2019, vol. 932, pp. 379–390, https://doi.org/10.1007/978-981-13-6052-7_33.

[25] P. Danenas, T. Skersys, and R. Butleris, "Natural language processing-enhanced extraction of SBVR business vocabularies and business rules from UML use case diagrams," *Data Knowl. Eng.*, vol. 128, no. June 2019, p. 101822, 2020, https://doi.org/10.1016/j.datak.2020.101822.

[26] B. Estrada-Torres, M. Camargo, M. Dumas, L. García-Bañuelos, I. Mahdy, and M. Yerokhin, "Discovering business process simulation models in the presence of multitasking and availability constraints," *Data Knowl. Eng.*, vol. 134, no. December 2020, p. 101897, 2021, https://doi.org/10.1016/j.datak.2021.101897.

[27] T. Skersys, P. Danenas, R. Butleris, A. Ostreika, and J. Ceponis, "Extracting sbvr business vocabularies from uml use case models using m2m transformations based on drag-and-drop actions," *Appl. Sci.*, vol. 11, no. 14, 2021, https://doi.org/10.3390/app11146464.

[28] Q. Zeng, H. Duan, and C. Liu, "Top-Down Process Mining from Multi-Source Running Logs Based on Refinement of Petri Nets," *IEEE Access*, vol. 8, pp. 61355–61369, 2020, https://doi.org/10.1109/ACCESS.2020.2984057.

[29] Q. Mo, F. Dai, D. Liu, J. Qin, Z. Xie, and T. Li, "Development of private processes: A refinement approach," *IEEE Access*, vol. 7, pp. 31517–31534, 2019, https://doi.org/10.1109/ACCESS.2018.2889715.

[30] E. Bazhenova, F. Zerbato, B. Oliboni, and M. Weske, "From BPMN process models to DMN decision models," *Inf. Syst.*, vol. 83, pp. 69–88, 2019, https://doi.org/10.1016/j.is.2019.02.001.

[31] J. T. Licardo, N. Tankovic, and D. Etinger, "A Method for Extracting BPMN Models from Textual Descriptions Using Natural Language Processing," in *Procedia Computer Science*, 2024, vol. 239, no. January, pp. 483–490, https://doi.org/ 10.1016/j.procs.2024.06.196.

[32] S. Sholiq, R. Sarno, and E. S. Astuti, "Generating BPMN diagram from textual requirements," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, pp. 10079–10093, 2022, https://doi.org/ 10.1016/j.jksuci.2022.10.007.

[33] S. Sholiq, M. A. Yaqin, A. P. Subriadi, and B. Setiawan, "Generation of business process modeling notation diagrams from textual functional requirements in Indonesian," *Int. J. Electr. Comput. Eng.*, vol. 15, no. 3, pp. 2938–2950, 2025, https://doi.org/10.11591/ijece.v15i3.pp2938-2950.

[34] X. Han *et al.*, "A-BPS: Automatic Business Process Discovery Service using Ordered Neurons LSTM," *Proc. - 2020 IEEE 13th Int. Conf. Web Serv. ICWS 2020*, vol. 1, pp. 428–432, 2020, https://doi.org/10.1109/ICWS49710.2020.00063.

[35] A. Mustansir, K. Shahzad, and M. K. Malik, "Sentiment Analysis of User Feedback on Business Processes," in *Proceedings - 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops, ASEW 2021*, 2021, pp. 204–210, https://doi.org/https://doi.org/10.1109/ASEW52652.2021.00048.

[36] S. Ahmed and K. Shahzad, "Augmenting Business Process Model Elements With End-User Feedback," *IEEE Access*, vol. 10, no. October, pp. 115635–115651, 2022, https://doi.org/10.1109/ACCESS.2022.3216418.

[37] K. Baïna, M. El Hamlaoui, and H. Kabbaj, "Business Process Modelling Augmented: Model Driven transformation of User Stories to Processes," *ACM Int. Conf. Proceeding Ser.*, pp. 215–220, 2020, https://doi.org/10.1145/3419604.3419793.

[38] S. Ahmed and A. Mustansir, "Mapping Textual Feedback to Process Model Elements," *Proc. - 2020 35th IEEE/ACM Int. Conf. Autom. Softw. Eng. Work. ASEW 2020*, pp. 131–137, 2020, https://doi.org/10.1145/3417113.3423376.

[39] S. Nasiri, Y. Rhazali, M. Lahmer, and N. Chenfour, "Towards a Generation of Class Diagram from User Stories in Agile Methods," in *Procedia Computer Science*, 2020, vol. 170, pp. 831–837, https://doi.org/10.1016/j.procs.2020.03.148.

[40] D. C. Wynn and A. M. Maier, "Feedback systems in the design and development process," *Res. Eng. Des.*, vol. 33, no. 3, pp. 273–306, 2022, https://doi.org/10.1007/s00163-022-00386-z.

[41] L. Yan *et al.*, "Practical and ethical challenges of large language models in education: A systematic scoping review," *Br. J. Educ. Technol.*, vol. 55, no. 1, pp. 90–112, 2024, https://doi.org/10.1111/bjet.13370.

[42] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Human Behav.*, vol. 104, 2020, https://doi.org/10.1016/j.chb.2019.106189.

[43] A. Praveen Kumar, A. Nayak, K. Manjula Shenoy, R. J. Manoj, and A. Priyadarshi, "Pattern-Based Syntactic Simplification of Compound and Complex Sentences," *IEEE Access*, vol. 10, pp. 53290–53306, 2022, https://doi.org/10.1109/ACCESS.2022.3174846.

[44] A. Haj, A. Jarrar, Y. Balouki, and T. Gadir, "The Semantic of Business Vocabulary and Business Rules: An Automatic Generation from Textual Statements," *IEEE Access*, vol. 9, pp. 56506–56522, 2021, https://doi.org/10.1109/ACCESS.2021.3071623.

[45] M. S. M. Rudwan and J. V. Fonou-Dombeu, "Hybridizing Fuzzy String Matching and Machine Learning for Improved Ontology Alignment," *Futur. Internet*, vol. 15, no. 7, 2023, https://doi.org/10.3390/fi15070229.

[46] S. Van Woensel, W.; Motie, "NLP4PBM: A systematic review on process extraction using natural language processing with rule based, machine and deep learning methods," *Enterp. Inf. Syst.*, 2024, https://doi.org/10.1080/17517575.2024.2417404

[47] X. Qin, S. Li, and H. Liu, "Adaptive fuzzy synchronization of uncertain fractional-order chaotic systems with different structures and time-delays," *Adv. Differ. Equations*, vol. 2019, no. 1, pp. 1–16, 2019, https://doi.org/10.1186/s13662-019-2117-1.

Grey Wolf Optimization-Based Decision-Control for Multi-Robot Signal Source Localization in Communication-Constrained Environments

Xuanwei Zhang
College of Information Science and Technology of Beijing University of Chemical Technology, Beijing, 102200, Beijing, China
E-mail: zxwzxw213213@163.com

Keywords: signal source localization; multi-robot system; Hazardous Environment; consensus control; GWO

Received: June 30, 2025

By developing and evaluating a decision-control strategy using a Hazardous Environment method, this study addresses the issue of signal source localization utilizing a group of autonomous robots. A new grouping technique is the basis of our proposed algorithm, Optimal Weighting Grey Wolf Optimization. The OW-GWO algorithm takes into account the past, present, and future ideal positions of all grey wolves, ranks them according to these positions, and updates these positions as needed. The alpha wolf constantly estimates where the prey is, and the rest of the grey wolves follow suit. Integrating the OWGWO method with an enhanced grouping strategy and dividing the algorithm into two stages—the random walk stage with the dynamic grouping stage—allows us to address the multi-target issue of swarm robots search. In the random walk phase, grey wolves adjust their best placements based on history and travel at random. The OWGWO algorithm builds search auxiliary points throughout the dynamic grouping stage by using a better grouping approach that takes into account the previous ideal placements of individual grey wolves. In order to hunt for various prey, grey wolves use these to form groups. . A decision level along with a control level is both included in the proposed decision-control method. The decision level employs a particle filter to make educated guesses about where the signal could be coming from. The actual location of the signal source becomes closer to the predicted position as the robots move. At the level of control, a consensus controller is suggested for commanding several robots to locate a signal source according to the predicted location of the source. Additionally, in order to alleviate some of the communication load, a Hazardous Environment mechanism is developed. Lastly, experiments and simulations demonstrate that the suggested decision-control method using the Hazardous Environment strategy solves the signal source localization issue well.

Povzetek: Prispevek predstavi strategijo vodenja skupine avtonomnih robotov za lokalizacijo signalnega vira v nevarnem okolju, ki z optimizacijskim algoritmom OW-GWO in dinamičnim grupiranjem izboljša učinkovitost iskanja več ciljev.

1 Introduction

The signal source localization issue has attracted a lot of interest for multi-robot systems as an alternative to the single robot, due to their vast detection range with simultaneous sampling. One typical approach to estimating the location of the signal source is the combined gradient estimation of the signal strength distribution. This method involves directing the movement of multiple robots to take measurements at different locations at the same time, allowing for the neglect of some unnecessary trajectories [1]-[4]. As an example, we suggested distributed control techniques for locating a noisy signal source and used a Finite-Difference approach to estimate the signal intensity gradient at the formation centroid. In, the search environment was modeled by Lu using a radial basis function network. Then, using gradient information supplied by the model, the robots were instructed to approach the signal source. Similarly, consensus control theory has informed the development of some cooperative control systems. Also, crisis management is an area where the concept of shared control is expanded upon. As an example, in, Garca-Magariño put out a method for citizen cooperation to identify issue causes via the utilization of a global map and peer-to-peer communication. The aforementioned methods, however, need longer measurement collection times from the robot at various sites.

Additionally, it is often unnecessary for these methods to provide certain search trajectories [5]-[9].

Two potential problems with the previously suggested methods for solving the signal source localization problem need to be highlighted [10]. One problem is that local optima may be readily achieved using the gradient estimation approach because to its sensitivity to noise. The uncertainty problem caused by noise may be addressed using a particle filter technique. Another problem with multi-robot systems Nomenclature is that there is a lack of available communication capacity for each robot. This problem may be solved by implementing an event-triggered strategy that shortens the time it takes for each robot to communicate [11]. Noteworthy among the suggested rules for multi-robot systems are those that are activated by events. Having said that, event-triggered rules like this primarily serve to save computing resources. Maintaining system stability in multi-robot systems still requires continual communication mechanisms [12]-[17].

Symbol / Term	Description
SLAM	A method that allows robots to simultaneously estimate their location and map unfamiliar areas is known as simultaneous localization and mapping.
DRL	Deep Reinforcement Learning is a method of learning that maximizes control policies by use of exploration motivated by rewards.
LSTM	The Long Short-Term Memory model is a recurrent neural network that can adapt to a changing environment and forecast temporal sequences.
GWO	The Grey Wolf Optimizer is a metaheuristic algorithm that takes its cues from the natural world, namely the hunting behavior and social structure of grey wolves.
OW-GWO	Fair Weighting An enhanced version of Grey Wolf Optimization, GWO incorporates a weighted rating of wolves according to their current, historical, and prospective positions.
α (Alpha) Wolf	Wolf pack leader in GWO who directs the pack's quest for the best possible location to catch prey.
β (Beta) Wolf	Helping the alpha with decision-making and search refining as a second-level wolf.
δ (Delta) Wolf	A wolf of the third rank who follows α and β and leads the pack.
ω (Omega) Wolves	Following their leaders, the surviving wolves scour the search area for the best possible answers.
RMSE	A statistical metric for assessing the precision of signal source localization predictions is the Root Mean Square Error.
SR (%)	Rate of Success—the proportion of times the multi-robot system was able to successfully localize its target.
RE (%)	The term "route efficiency" refers to the ratio between the ideal and actual lengths of paths that robots follow.
CT (s)	The merging of the amount of time it takes for the swarm to get to the starting point.
CO (MB/s)	The average amount of data exchanged per second between agents operating in a swarm is known as communication overhead.
PF	The decision layer uses the particle filter, a probabilistic estimating method, to anticipate the location of the source.
CE	Distributed control technique that guarantees consistent agreement among robots in decisions is the consensus estimator.
HE-DRL	Potentially Dangerous Setting Deep Reinforcement Learning—a kind of DRL designed to deal with ambiguity and incomplete data.
V-SLAM	Visual SLAM is a kind of SLAM that maps and perceives its surroundings using visual sensors, such as cameras.
$\hat{x}(t)$	This is the predicted location of the signal at time t.
$x(t)$	True location of the signal source at time t.
Δx	Localization error among estimated and real locations.
η	Acceleration of learning or factor for adaptability in DRL model.
γ	Reinforcement learning discount factor as a subsequent reward.

ρ	Issue of communication dependability in potentially dangerous environments.
N_a	Amount of swarm-based autonomous robots.
d_{max}	What is the maximum distance that agents may communicate?
P_{tot}	The overall amount of electricity used by the swarm system while it is operating.

A. Contributions

A number of Hazardous Environment systems have been developed to save communication resources and lessen the computational and communication loads. Unfortunately, an answer is yet unavailable for the signal source localization issue that combines a Hazardous Environment system with a cooperative control technique and particle filter methodology. The provided cooperative control strategy presents a number of challenges, one of which is the design of Hazardous Environment rules. The second obstacle is figuring out how to use the Hazardous Environment rule to derive stability criteria for multi-robot systems using the cooperative control technique that has been suggested. Thus, the current work is motivated by the need to address how to create the decision-control technique for the signal source localization issue while considering the aforementioned constraints. Two benefits characterize the suggested decision-control method. The implementation of a Hazardous Environment strategy has the potential benefit of reducing communication delays and the frequency of control input updates, which in turn saves communication and chip resources. Additionally, the multi-robot system's detection data may be effectively used through the particle filter with cooperative controller to approximate the signal source's location.

B. Novelty and technical innovations

An innovative approach to multi-robot search has been developed by combining state-of-the-art optimization, control, and machine learning methods:

- OW-GWO, or Optimal Weighting Grey Wolf Optimization, is a novel writers' method for grouping. It differs from conventional GWO in that it takes into account the ideal positions of the "grey wolves" (robots) in the past, present, and future while ranking and updating their positions. The multi-robot search technique relies on this improvement.
- There are a random walk phase and a dynamic grouping phase in a two-stage search algorithm. Swarm robot search's multi-target problem may be tackled in a new manner using this structured method and the OW-GWO.
- Hazardous Environment Mechanism (DRL using LSTM): A new way to reduce robots' communication burden is to employ a Deep Reinforcement Learning (DRL) technique combined by a Long Short-Term Memory (LSTM) network. This is very important for scalability and real-time performance, particularly in difficult circumstances.
- The suggested approach divides the system into two main levels: the Decision Level, which uses a Particle Filter to estimate the source, and the Control Level, which uses a Consensus Controller to navigate. This architecture allows for integrated decision-control processing. Together, these distinct features offer a solid basis for concerted effort.

C. Key findings of the study

The following important conclusions were shown by the research, which confirmed the efficacy of the suggested system via tests and simulations:

- Successful Signal Source Localization: The proposed decision-control approach successfully resolves the signal source localization problem.
- Improved Convergence: The real signal source location becomes closer to the anticipated position as the robot’s advance, thanks to the Particle Filter at the decision level, which improved convergence.
- Coordinated Control: The many robots were effectively directed by the Consensus Controller to find the signal source using the projected position, showcasing efficient coordinated movement.
- Communication Efficiency: An essential discovery for developing scalable multi-robot systems is the improvement in communication efficiency, which was intended to be a byproduct of incorporating the Hazardous Environment method (DRL with LSTM).

In order to assess the suggested Decision-Control Strategy utilizing the Hazardous Environment (DRL + LSTM) approach along with Optimal Weighting Grey Wolf Optimization (OW-GWO) for signal source

identification under visual SLAM, here is a numerical comparison summary of the study you mentioned, with an emphasis on the success rate, route efficiency, and RMSE performance metrics.

D. Paper structure

Here is how the rest of the article is structured. The second section provides a concise overview of the basics of the communication topologies and dynamics of mobile robots. Section 3 detail our plan to coordinate the movement of mobile robots in order to pinpoint the location of the signal source using a particle filter-based position estimation and a Hazardous Environment mechanism. We shall demonstrate the efficacy of the suggested decision-control method using the Hazardous Environment system in Section 4 via simulation and in Section 5 through experimental data. Section 6 will serve as the paper's conclusion.

2 Related work

In order to increase the effectiveness of odor source search while preventing collisions, researchers from [21] suggested a distributed behavior control strategy for indoor multi-robot collaborative odor source localization. For situations like drug screening, fire rescue, and others like it, that challenge is critical. In a spiraling motion, a number of robots seek for the plume. By exchanging data on the spot with the greatest concentration of detected odors, they may speed up the system's plume detection process by directing gas-detecting robots to that spot. In order to pinpoint where the smell is coming from, a number of robots use particle filters. They determine where the smell is coming from by establishing limits for particle convergence with distance convergence. An efficient odor source search may be accomplished with the help of a state machine, which allows for the flexible selection of the most suitable approach based on the present environment and job requirements.

A distributed cooperative navigation technique is presented by the authors of [22], who overcome these restrictions. The first step is to apply a Distributed Adaptive Extended Kalman Filter. That filter can handle multi-source noise and nonlinearities by adaptively adjusting the noise covariance matrix. Later on, a framework called Distributed Model Predictive Control is presented. By optimizing and predicting the kinematic models of each robot, that framework enhances the system's capacity for joint operations and dynamic decision-making. Robot tests and comprehensive simulations validate the strategy's effectiveness. Simulation findings show that DAEKF achieves better localization accuracy than Kalman Filter with Extended Kalman Filter when it comes to cooperative localization. Using DAEKF, the two robots significantly decreased their lateral and heading errors in the straight-line path-tracking trials. The lateral error Root Mean Squared Error for Robot 1 was decreased by 68.87%, 27.80%, and 25.76% when using DAEKF instead of No Filtering, KF, or EKF, respectively. By using DAEKF, the RMSE was decreased with 52.29%, 41.89%, along with 36.47% in the heading error case. In comparison to No Filtering, KF, and EKF, DAEKF decreased the lateral error RMSE for Robot 2 through 51.30%, 22.88%, & 11.60%, respectively.

A human-operated multi-robot system for source localization was described in [23] by the authors. The system included both ground and airborne robots with varying degrees of autonomy. The technology allows a human supervisor to manage several robots via augmented reality interfaces. That human-robot interface's primary function is to facilitate the cooperative, real-time control of diverse robot groupings. To free up the operators for more complex duties, it avoided obstructions using sophisticated route planning algorithms. The surroundings and barriers are known to each robot, so it can autonomously construct a route to any user-selected objective without any collisions. In the video view, it showed sensor data from each robot individually. Additionally, an augmented reality perspective that incorporates sensor data is shown, which assists users in locating information sources or the operator in accomplishing mission objectives. That study delves into an early Human Factors assessment of the system, testing various interface circumstances for source task identification. Mission completion durations for target detection missions were shown to be lower using the innovative Augmented Reality multi-robot control (Point-and-Go with Path Planning) as compared to the standard joystick control. Additionally, they look at operator workload analysis and usability testing.

Researchers in [24] demonstrated a human-operated multi-robot system that could locate sources with relative ease using ground and airborne robots with some degree of autonomy. The technology allows a human supervisor to manage several robots via augmented reality interfaces. That human-robot interface's primary function is to facilitate the cooperative, real-time control of diverse robot groupings. To free up the operators for more complex duties, it avoided obstructions using sophisticated route planning algorithms. The surroundings and barriers are known to each robot, so it can autonomously construct a route to any user-selected objective without any collisions. In the video view, it showed sensor data from each robot individually. Additionally, an augmented reality perspective that incorporates sensor data is shown, which assists users in locating

information sources or the operator in accomplishing mission objectives. That study delves into an early Human Factors assessment of the system, testing various interface circumstances for source task identification. Mission completion durations for target detection missions were shown to be lower using the innovative Augmented Reality multi-robot control (Point-and-Go with Path Planning) as compared to the standard joystick control. Additionally, they look at operator workload analysis and usability testing.

Current techniques' authors rely heavily on responsive stimuli or simple gas plume models to plan robot movements [25]. The patchy distribution of real-world gases makes these methods ineffective, even if they work well in idealized, simulated systems. They present SniffySquad, an olfactory-based multi-robot system, to tackle the intrinsic patchiness in gas source location. To improve data gathering and estimate, SniffySquad uses an active sensing technique that takes patchiness into account. In addition, it boosts effectiveness of source-seeking initiatives using a unique collaborative role adaption technique. The results of their system's thorough assessments show that it outperforms state-of-the-art gas source localization systems, increasing the success rate by 20%+ and improving the route efficiency by 30%+.

The authors of [26] presented Swarm-SLAM, a free and open-source C-SLAM system with the characteristics essential to swarm robotics: scalability, flexibility, decentralization, and sparsity. In order to decrease communication and speed up convergence, their system incorporates a unique inter-robot loop closure prioritizing method and supports lidar, stereo, as well as RGB-D sensing. They tested their ROS-2 setup in a real-world setting with three ad hoc-networked robots and on five separate datasets.

The primary methods of multi-robot collaborative SLAM were presented by the researchers in [27], along with the concepts and typical approaches of visual SLAM. In that study, they took a look at the three primary issues plaguing multi-robot collaborative visual SLAM systems today: back-end optimization, map fusion, and job allocation. Afterwards, that study outlined several alternatives and evaluated their pros and cons. That work also aims to point future research in related domains in the right direction by introducing some potential future research directions of multi-robot collaborative visual SLAM technology.

Before proposing their new searching algorithm, Entrotaxis-Turn, the authors of [28] combine a cognitive search strategy (the Entrotaxis method) with an intermittent search strategy, which has remarkable theoretical performance. In order to improve the success rate and save searching time, the hybrid algorithm may use the triggered turn motion to prevent slipping into a loop by avoiding obstacles. Secondly, they expand the ET method to the Multi-ET method, which allows several robots to work together and exchange data acquired by each mobile sensor. In order to demonstrate how well the algorithms work, a standard chemical cluster scenario is utilized to build a simulation and diffused gas is created using the advection-diffusion equation. they find the best ET and Multi-ET parameter combinations using Monte Carlo simulations. Lastly, the computational fluid dynamics model's simulation scenario is used to conduct a verification experiment. The results demonstrate that the ET algorithm is far more efficient and has a much higher success rate than the Entrotaxis method. In addition, the Multi-ET algorithm achieves the best possible performance while looking for sources.

A multi-modal, multi-robot environmental sensing technique that is customized to dynamic real-world situations was presented by the authors of [29] as a state-machine model. In order to localize and map gas sources, the multi-modal algorithm combines two different exploration strategies: first, an initial exploration phase that uses variable formations and multi-robot coverage path planning to provide early gas field indication; while second, an active sensing phase that uses multi-robot swarms to provide precise field estimation. The mechanism for switching among these two sensing methods is provided by the state machine. While in the phase of exploration, a coverage route is created that maximizes the area that is explored, and at pre-defined sample periods, it measures the concentration of gas and estimates the first gas field. Afterwards, during the active sensing phase, a multi-robot swarm system's coordinated and effective sensing is ensured as mobile robots operating in a swarm work together to choose the next measurement location by broadcasting possible positions and reward values. The process of validating the system includes doing tests using hardware-in-the-loop and real-time trials using a radio source that mimics a gas field. Modern gas source localization and state-of-the-art single-mode active sensing methods are used to thoroughly evaluate the suggested method.

To tackle the issue of source localization and multi-robot area coverage in unfamiliar contexts, the authors of [30] presented a distributed online learning coverage control framework that is based on sparse Gaussian process regression. That study uses multiple robots to explore the task area and gather environmental information. It then uses variational free energy methods to determine the posterior distribution of the model, which is input to the centroid Voronoi

tessellation algorithm. That approach overcomes the constraints of traditional Gaussian process regression when dealing with large datasets. Furthermore, in order to achieve autonomous online decision-making while optimal coverage, algorithms for centroid Voronoi tessellation with separating hyperplanes and buffer factors were developed for dynamic robot work area planning, which takes into account the localization errors and the impact of obstacles. The simulation results show that the suggested method enhances source localization accuracy, increases iteration speed, and guarantees the safety of multi-robot structures, demonstrating the usefulness of the model changes. Table.1. shows the Multi-Robot System Strategy Approaches of Comparison.

Table 1: Multi-robot system strategy approaches comparison

Reference	Technique / System	Primary Application / Goal	Key Technical Approaches	Limitations (Stated or Inferred)
[21]	State machine and spiraling distributed behavior control	Location of Indoor Odor Sources (for instance, in drug screening and fire rescue operations).	Various methods for preventing collisions, include spiral searches, particle filters, and state machines for approach selection flexibility.	Assumed: Particle filter convergence and unambiguous plume detection are crucial for performance.
[22]	DAEKF and DMPC Framework	Accurate path-tracking and distributed cooperative navigation.	For handling noise and non-linearities, we have the Distributed Adaptive Extended Kalman Filter (DAEKF). For generating dynamic decisions, we have the Distributed Model Predictive Control (DMPC).	Inference: DMPC optimization may need a significant amount of computer resources.
[23] / [24]	State machine and spiraling distributed behavior control	Various ground and aerial robots may be used for source localization with target identification.	Supervised autonomous obstacle avoidance using an Augmented Reality (AR) interface and intelligent route planning.	The following is inferred: the need for a human supervisor (possible bottleneck in the workload); the need for course planning based on known obstacles and environments.
[25]	DAEKF and DMPC Framework	Localizing Gas Sources by Addressing Real-World Gas "patchiness."	Active sensing using patchiness; innovative collaborative role adaption method.	Tackles the shortcomings of straightforward plume models, resulting in a more intricate system.
[26]	State machine and spiraling distributed behavior control	Swarm robots using collaborative SLAM (C-SLAM).	An approach for prioritizing the closure of inter-robot loops that is scalable, decentralized, and supports lidar, stereo, and RGB-D.	Inferred: Three robots were used for real-world validation.
[27]	DAEKF and DMPC Framework	Assessment of Multi-Robot Collaborative SLAM.	Examines the fundamentals of visual SLAM, optimization of the backend, fusion of maps, and task allocation.	The main constraints in this area are highlighted in this paper: optimization of the back-end, fusion of

				maps, and allocation of jobs.
[28]	State machine and spiraling distributed behavior control	Efficiently Locating the Origin of Odors...	A combination of cognitive (Entrotaxis) as well as intermittent search tactics; action to avoid loops and barriers via triggered rotation motion.	Inferred: Verification by computational fluid dynamics (CFD) models and the advection-diffusion equation.
[29]	DAEKF and DMPC Framework	Finding and Charting Gas Sources in Real-Time Environments.	The two parts of a two-stage state machine are 1) exploration-based coverage route planning and 2) accurate estimation-based swarm-based active sensing.	Presumably, rather of using real gas diffusion, real-time validation made use of a radio source to simulate a gas field.
[30]	State machine and spiraling distributed behavior control	When working in new environments, source localization with area coverage is crucial.	Dynamic work area planning using Centroid Voronoi Tessellation (CVT) and Sparse Gaussian Process Regression (GPR) for handling big datasets.	Validation derived from simulation outcomes; GPR complexity (including sparse GPR) inferred.

3 Problem statement

A collection of robots working together to explore and map an unfamiliar area is called a multi-robot collaborative autonomous exploration task. Using sensors like LiDAR, each robot scans its environment and builds its own customized map as part of the exploration operation. On a regular basis, the bots will provide the server their geolocation data. Each robot is able to choose a target site due to the exploration strategy's decentralized implementation. The route planning algorithm simultaneously guides the robot to the desired destination as it moves, enabling it to explore its surroundings. Two robots may exchange local maps, positions, along with target locations if their distance is less than the communication distance d_c. Whenever a robot completes its excursion, the server instantly incorporates its map into the larger map. The exploration job is completed when the server gives an end command to every robot, marking the completion of the global map.

In order to reduce interior air pollution and ensure the safety of those within, source localization is crucial. In a Human-Robot Collaboration setting, for instance, robots would have to do safety analyses to ensure they stay out of harm's way. The majority of research has been on static sources and mechanical ventilation, but this work delves into the less-explored area of finding sources that fluctuate over time in naturally ventilated areas. For warehouses to be safe and efficient places to work, there are a number of risks associated with using multi-robot systems. The main causes of these dangers are the difficulties in controlling many robots in tight areas and the interactions between robots with human employees. To improve upon the limitations of conventional 2D systems that are based on a set height, we have created a novel 3D localization method that can adapt to different heights. We compared 2D and 3D approaches, simulating natural ventilation using a swinging fan, to guarantee uniformity in the experimental setup. Overall, warehouse operations are made safer with the help of autonomous robots built for safety and health checks. These robots can identify risks on their own and inform other systems and human workers in real-time.

We think about a setup with K robots and an RF source that is both stationary and far away. Robot k's location during the nth control cycle is represented by $\mathbf{x}_{n,k} = \left[x_{n,k}, y_{n,k}\right]^T$, besides the source is situated at $\mathbf{x}_{n,0} = [x_0, y_0]^T$ with $n = 1, \cdots, N$. The robots' mission is to navigate themselves to the source. Since we don't know where the source is or where the robots are, we'll employ a combined probabilistic model to try to guess where they are all at once. Let us pretend the robots are able to pick up reference RF signals that they may use to gauge their relative positions and the distances

between themselves and one another. The following distance measuring model is used in this study:

$$p\left(z_{n,k_1,k_2} \mid \mathbf{x}_{n,k_1}, \mathbf{x}_{n,k_2}\right) = \mathcal{N}\left(z_{n,k_1,k_2}; \alpha_0 + r_{n,k_1,k_2}\alpha, r_{n,k_1,k_2}\sigma_z^2\right) \quad (1)$$

where z_{n,k_1,k_2} stands for the robot's distance measurement as measured by the RF signal it received $k_2(k_2 = 1, \cdots, K)$ besides transmitted from robot $k_1(k_1 = 1, \cdots, K)$ otherwise the source $(k_1 = 0)$. $\mathcal{N}(z; \mu, \sigma^2)$ (z) is a Gaussian random variable with a mean of μ and a variance of μ, and its probability density function is σ^2. $r_{n,k_1,k_2} = \left\|\mathbf{x}_{n,k_1} - \mathbf{x}_{n,k_2}\right\|_2$ is the distance among node k_1 and node k_2. α_0, α, and σ_z^2 describe characteristics of the actual physical setting.

The robots must choose their next move based on the measured distance in each control cycle and then create control inputs to implement that decision. In two successive control cycles, the robot's locations are defined as

$$\mathbf{x}_{n,k} = \mathbf{x}_{n-1,k} + \mathbf{c}_{n-1,k} + \mathbf{n}_{n-1,k} \quad (2)$$

where $\mathbf{c}_{n-1,k}$ During the $\mathbf{n}_{n-1,}$ th control cycle, is the input for robot k's control, and $\mathbf{n}_{n-1,k}$ is a controlling error that follows a normal distribution with a mean of zero and a variance of squared. Because of this, we can calculate the distribution of the robot's location throughout transit as

$$p\left(\mathbf{x}_{n,k} \mid \mathbf{x}_{n-1,k}\right) = \mathcal{N}\left(\mathbf{x}_{n,k}; \mathbf{x}_{n-1,k} + \mathbf{c}_{n-1,k}, \sigma_c^2\mathbf{I}\right) \quad (3)$$

where [I] is a matrix of identities. A further benefit of the source's stationary nature during the searching process is that its position transition distribution may be expressed as

$$p\left(\mathbf{x}_{n,0} \mid \mathbf{x}_{n-1,0}\right) = \delta\left(\mathbf{x}_{n,0} - \mathbf{x}_{n-1,0}\right) \quad (4)$$

given that δ(·) is an impulse function of Dirac type.

When just starting out, $p\left(\mathbf{x}_{1,0}\right) \triangleq p\left(\mathbf{x}_{1,0} \mid \mathbf{x}_{0,0}\right)$ should be distributed uniformly over the investigated region, and $p\left(\mathbf{x}_{1,k}\right) \triangleq p\left(\mathbf{x}_{1,k} \mid \mathbf{x}_{0,k}\right) = \delta\left(\mathbf{x}_{1,k} - \overline{\mathbf{x}}_{1,k}\right)$ for $k = 1, \cdots, K$, where $\overline{\mathbf{x}}_{1,k}$ represents robot k's actual starting point. The assumption that the robot locations are known at n=1 is not sufficient to deduce their positions when n≥2. Instead, one must consider both the starting positions with the control inputs. The reason for this is because if the control error $\mathbf{x}_{n,k}$ is left unchecked, it may build up over time and lead to very erroneous robot position predictions.

Here, given the locations of the robots, we craft the control strategy by making the most accurate predictions about the relationship between the source's position with the distance measurements. In order to adequately quantify the decrease in source position uncertainties when distance data are available in the following control cycle, we use mutual data as the objective function.

Minimizing the time of the exploration process is the purpose of the autonomous exploration approach. Hence, the following is the definition of the objective function:

$$min[\sum_{i=1}^{T} t_i], \quad (5)$$

such that t_i is the length of the ith time step and T is the total amount of time steps in the exploration phase.

4 Decentralized multi-robot model

A D-MDP, which stands for Decentralized Partially Observable Markov Decision Process, is used to represent the decentralized multi-robot autonomous exploration issue. The tuple is what defines a D-MDP $\langle I, S, \mathbb{A}, P, R, \mathbb{O}, O, h\rangle$. The tuple's components are the sets of robots (l), states (S), and finite sets of joint actions (A). The state transfer likelihood, $P(s' \in S, a)$, indicates the likelihood that, while in state s, the robot team performs the joint action a and thereafter moves to state s'. $s, s' \in S, a \in \mathbb{A}. R(s, a, s')$ signifies the reward that the robot team receives in state s after the execution of the joint action with an and subsequent transition to state s^', with $s, s' \in S$ and $a \in \mathbb{A}. \mathbb{O}$ is the finite set of joint observations. $O(s', z)$ is the function that, when the environment is in state s^', offers the likelihood to acquire the joint observation z, where $s' \in S$ and $z \in \mathbb{O}$. The area is to find a set of distributed policies $\Psi = \{\pi_i\}_{i\in I}$ that may allow the robotics squad to optimize their anticipated team benefits during the course of an episode's horizon h:

$$\Psi^* = \arg\max_{\psi} \mathbb{E}\left[\sum_{t=0}^{h-1} \gamma^t R(s_t, a_t, s_{t+1}) \mid s_0, \psi\right], \quad (6)$$

where $\gamma \in (0,1]$ is a discount feature for future rewards, s_0 is the primary state. Fig 1 (a) shows the overall architecture model.

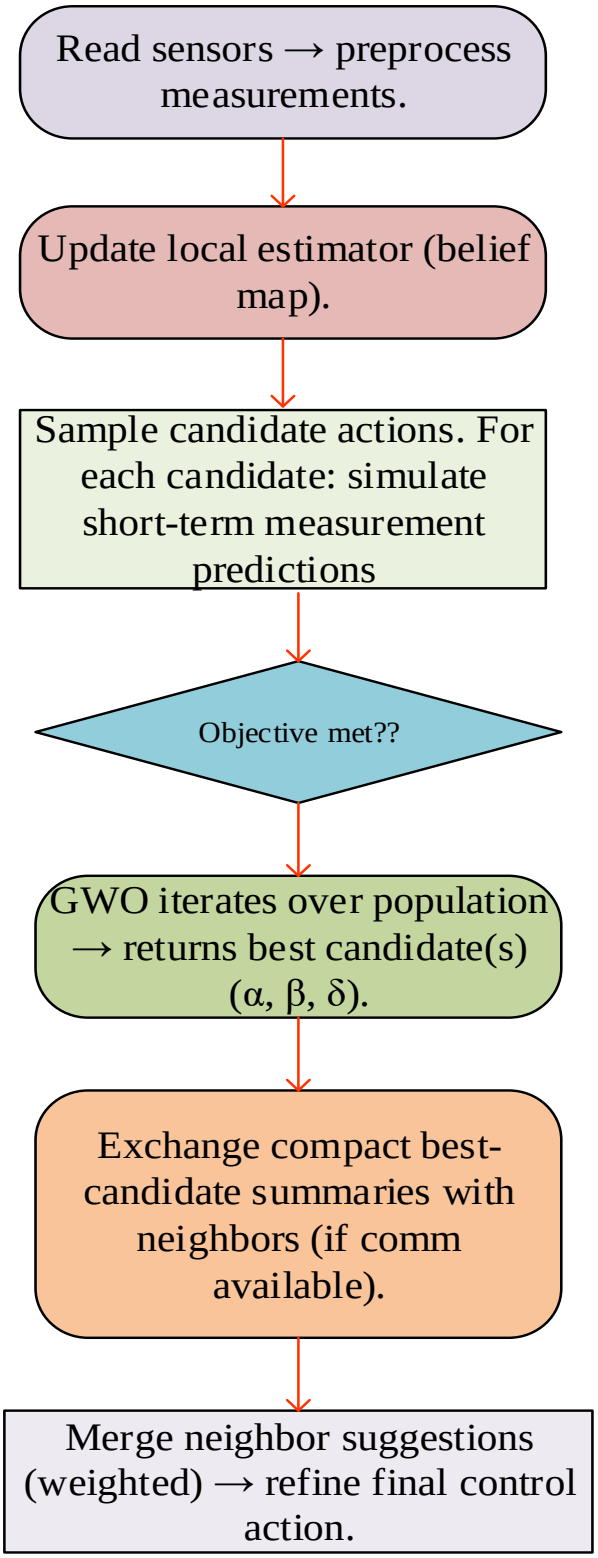


Figure 1: (a). Overall architecture

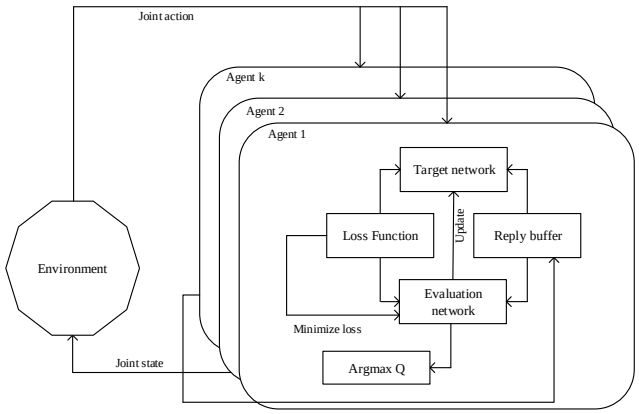


Figure 1: (b) Multi-Step DRL for Robot Localization

The first step in using DRL is deciding what you want to do in order to influence your surroundings. The robot is guided toward its destination via a low-level controller and a route planning algorithm. In order to explore, the robot examines its surroundings as it moves. At the same time, the environment relays the data obtained from each robot to the main training system. The training framework stores mini-batches for future use and keeps track of the actions, observations, along with rewards from each environment encounter in the experience's replay pool.

4.1 Dataset

- **Multi-Robot Warehouse information set:** With details like the robot's position, job specifications, and obstacle placements, the Multi-Robot Warehouse information set aims to mimic the cooperative functioning of several robots in a warehouse setting. We found that Multi-Robot Warehouse is a better model to research logistics warehousing robot collaboration than AI4 Logistics because it focuses more on modeling collaborative operations and job distribution among several robots. Due to the coordinated efforts of several robots, the data collection is modest in size but very well-balanced.
- **ROS-based Logistics Simulations data set:** Information like as sensor readings, robot states, and a map of the surrounding area are all part of the ROS-based Logistics Simulations data set. The simulation quality of this dataset makes it ideal for use in training and validating

deep learning models. The information set is reasonable in size and has properties that make it consistent and controllable due to its production in the ROS model environment.

- **RoboCup Logistics:** According to that, the data set known as the RoboCup Logistics League is derived from the RoboCup Logistics League which contains information on how well robots perform in logistics and warehousing jobs. Despite its modest size, the dataset presents a more difficult assessment platform because of the realistic competitive situation. This data collection contains useful real-world scenario data for researching multi-robot collaborative operation techniques, including task execution information and instances of collaborative operations.

4.2 Data processing

In order to train and assess our DRL using Filtering model, data is crucial to this research. Included in the data preparation method are the following essential steps:

- Data collection: We gathered optical, lidar, along with ultrasonic sensor data from large-scale warehousing and logistics settings. These records provide crucial details including the robot's location, ambient condition, perception data, and job specifications.
- Data cleaning: We cleaned the data after collection to get rid of any outliers or incorrect information. Filtering sensor data for noise, eliminating anomalies from motion trajectories, and fixing data inconsistencies are all part of this process.
- Data annotation: In order to aid with the target identification process, certain data must be tagged. To make sure the dataset was useful and of high quality, we annotated it using open-source tools.
- Data partitioning: In order to train, tune, and evaluate the model, we divide the dataset into three parts: training, validation, and test. The goal of data segmentation is to improve the model's ability to generalize.
- Data format conversion: Our data format conversions were based on the model's input needs so that they could be used with various deep learning models and algorithms. Images may need to be resized and data may need to be normalized as part of the pre-processing stages.

Following these procedures, we were able to get a cleaned and annotated dataset, which served as a solid foundation for our studies. These numbers represent the logistics warehousing robots' impressions of the data and tasks needed in various settings, When dealing with datasets that include tens, hundreds, or even thousands of characteristics, this manual technique may quickly become overwhelming. Preprocessing the experimental data collected from sensors is the first step in developing the algorithm to address this feature selection challenge. In order to employ artificial neural networks for reliable prediction, data must first be pre-processed. As an initial step in pre-processing, features are normalized or standardised to make sure they don't impact each other's contributions to the neural network's structure.

There are several approaches to carry out the normalization procedure. Another option is to use the relation in (7) to make sure that all the characteristics' (data's) range is from 0 to 1. Data with a mean of 0 with a standard deviation of 1 are pre-processed in a different way, known as z-scoring, according to the relation in equation (8).

$$\text{Normalization} = \frac{\text{Sensor Data (feature)} - \text{Min}}{\text{Max} - \text{Min}}$$
$$\text{Standardization } (z - \text{ scoring }) = \frac{\text{Sensor Data (feature)} - \text{Mean}}{\text{Standard deviation}} \quad (7)$$

Data that has been pre-processed using normalization and z-scoring is shown in Figure 1. In light of its shown accuracy in other research, zscoring is used in this investigation. In order to choose which, feature to feed into the main LSTM, a new algorithm is built using deep reinforcement learning once the data has been pre-processed. The feature that provides the most accurate forecast of the output data having the least amount of inaccuracy is chosen.

4.3 Filtering

Think about a model for a generic (nonlinear) signal that includes the following equations $x_k \in R^n$ and measurement $y_k \in R^p$ with additive, uncorrelated noise $w_k \in R^n, v_k \in R^p$ of known statistics at the time case $k = \{0,1,2,3 \ldots n\}$.

$$x_{k+1} = f(x_k) + w_k$$
$$y_k = h(x_k) + v_k \quad (8)$$

The vectors $f(x_k)$ and $h(x_k)$ that are non-linear functions of x_k and k. The initial state x_0 has a defined probability density distribution and is uncorrelated to the noise processes outlined before p_0. A cost function at time k that is two-parameter, risk-sensitive, and squared error type may be described as

$$J_{RS}(\zeta, k) =$$

$$E\left[exp\ \left(\mu_1 \sum_{i=0}^{k-1}\ (x_i - \hat{x}_i)^T(x_i - \hat{x}_i) + \mu_2(x_k - \varsigma)^T(x_k - \varsigma)\right)\right] \quad (9)$$

Where $\hat{x}_i$'s are the best guesses for the state variable's values from previous stages $i \in \{0,1,2,3 \ldots k-1\}$. The current optimum estimate $\hat{x}_k$ finds the best possible value of ϖ, which reduces $J_{RS}(\zeta, k)$,

$$\text{i.e. } \hat{x}_k = \underset{\varsigma}{arg\ min} J_{RS}(\varsigma, k) \quad (10)$$

The constant parameters μ_1 and μ_2 certain factors are known as risk-sensitive. The following recursive relations, established from prior papers, may be used to get the posterior solutions of the RSE issue.

$$\hat{x}_{k|k} = \underset{\varsigma}{arg\ min} \int\ exp\ [\mu_2(x_k - \varsigma)^T(x_k - \varsigma)]\alpha_k dx_k \quad (11)$$

where

$$\alpha_k = p(y_k \mid x_k) \times \int_{+\infty}\ exp\ |\mu_1(x_{k-1} - \hat{x}_{k-1})^T(x_{k-1} - \hat{x}_{k-1})|\alpha_{k-1}p(x_k \mid x_{k-1})dx_{k-1} \quad (12)$$

A posterior risk-sensitive particle filter is developed in this section.

A. Probabilistic interpretation

In order to get probability distributions, we normalize the distribution of $\{\alpha_k\}$ $\{\overline{\alpha_k}\}$ defined through

$$\bar{\alpha}_k(x_k) = \frac{\alpha_k(x_k)}{\int\ \alpha_k(x_k)dx_k} \quad (13)$$

Here is an alternative probability density function that we might consider.

$$\beta_{k-1}(x_{k-1}) = g(y_k \mid x_k)exp\ \left(\mu_1(x_{k-1} - \varsigma_{k-1})^T(x_{k-1} - \varsigma_{k-1})\right)\alpha_{k-1}(x_{k-1}) \quad (14)$$

From that probability density function, we may get the normalized version as

$$\bar{\beta}_{k-1}(x_{k-1}) = \frac{\beta_{k-1}(x_{k-1})}{\int\ \beta_{k-1}(x_{k-1})dx_{k-1}} \quad (15)$$

The normalised probability density $\{\overline{\alpha_k}\}$ uses the recently proposed probability density function to be expressed $\bar{\beta}_k$ as follows:

$$\overline{\alpha_k} = \frac{\int_{-\infty}^{+\infty} g(y_k|x_k)L\alpha_{k-1}p(x_k|x_{k-1})dx_{k-1}}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(y_k|x_k)L\alpha_{k-1}p(x_k|x_{k-1})dx_{k-1}dx_k} \quad (16)$$

Where $L = exp\ [\mu_1(x_{k-1} - \hat{x}_{k-1})^T(x_{k-1} - \hat{x}_{k-1})]$ (17)

4.4 Multi-robot SLAM with known initial poses

A solitary robot Assuming knowledge of the starting robot postures allows for easy extension of SLAM formalism to handle numerous robots. Think about two robots whose observations happen at the same exact moment: $x_{1:t}^1$ stands for the path that robot 1 will take, while x_(1:t)^2 represents the path that robot 2 will take in the opposite direction. Estimating the posterior probability across one map and two robot paths at the same time is our goal. We express this mathematically using factored form as:

$$p(x_{1:t}^1, x_{1:t}^2, m \mid z_{1:t}^1, u_{0:t-1}^1, x_0^1, z_{1:t}^2, u_{0:t-1}^2, x_0^2) = p(m \mid x_{1:t}^1, z_{1:t}^1, x_{1:t}^2, z_{1:t}^2)\ p(x_{1:t}^1 \mid z_{1:t}^1, u_{0:t-1}^1, x_0^1)p(x_{1:t}^2 \mid z_{1:t}^2, u_{0:t-1}^2, x_0^2) \quad (18)$$

in which the distribution over maps is given by the first term and the distribution over probable robot trajectories is given by the further terms. It is essential to note that this factorization is based on the assumption that both trajectories are separate and that the position of one robot does not affect the other's recorded observations (refer to Figure 2). We can design this assumption to be valid in the case of cooperative multi-robot mapping, even if it won't be true in general. Robots cannot rely on one another when they are far away (outside of sensor range or line-of-sight), but they can detect each other and ignore each other's observations when they are close by (within sensor range or line-of-sight). For instance, when the robots use laser range-finders, the subset of laser beams that hit other robots may be disregarded and mutual detection made easier by retro-reflective target.

The following is the construction of the particle filter for multi-robot SLAM. A particle's characteristics $\left\langle x_t^{1(i)}, x_t^{2(i)}, m_t^{(i)}, w_t^{(i)} \right\rangle$, where $x_t^{1(i)}, x_t^{2(i)}$ are the instantaneous poses for robots 1 and 2 , besides $m_t^{(i)}$ is the common map. Given an observation tuple $(z_t^1, u_{t-1}^1, z_t^2, u_{t-1}^2)$, Below is the updated filter step:

$$\begin{aligned} x_t^{1(i)} &= A\left(u_{t-1}^1, x_{t-1}^{1(i)}\right) x_t^{2(i)} = A\left(u_{t-1}^2, x_{t-1}^{2(i)}\right) \\ m_t^{(i)} &= M\left(z_t^1, x_t^{1(i)}\right) + M\left(z_t^2, x_t^{2(i)}\right) + m_{t-1}^{(i)} \\ w_t^{(i)} &= S\left(z_t^1, x_t^{1(i)}, m_{t-1}^{(i)}\right) S\left(z_t^2, x_t^{2(i)}, m_{t-1}^{(i)}\right) w_{t-1}^{(i)} \end{aligned} \tag{19}$$

just as in the case of a single robot, with the same action framework A, sensor model S, and map generator M. The outcome is easily applicable to a wide range of robots. There are two major caveats to this strategy from a practical perspective. One thing to keep in mind is that the state space will inevitably be bigger than in the case of a single robot, which means that more particles will be needed to achieve convergence. So, the state space for two robots that go 50 m each has the same dimension as the state space for one robot that travels 100 m, since the dimensionality grows with total route length rather than the total amount of robots. Therefore, given an effective exploration method (one in which the robots investigated zones overlap very rarely), a tiny particle set may provide excellent results. Secondly, if one robot stops while the other keeps moving, resampling might cause particle impoverishment near the robot that stopped, which is a restriction of the filter resampling process. Since filter divergence might occur due to poverty, it is optimal for all robots to move at the same pace.

4.5 Dynamics for mobile robots

The dynamics may be stated for mobile robots l in Figure 1 by:

$$\begin{pmatrix} \dot{r}_{xl} \\ \dot{r}_{yl} \\ \dot{\theta}_l \\ \dot{v}_t \\ \dot{\omega}_l \end{pmatrix} = \begin{pmatrix} v_l \cos\theta_l \\ v_l \sin\theta_l \\ \omega_l \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \frac{1}{m_t} & 0 \\ 0 & \frac{1}{h} \end{pmatrix} \begin{pmatrix} F_l \\ \tau_l \end{pmatrix} \tag{20}$$

where $r_i = \left(r_{xi}, r_{yi}\right)^T$ is the situation of the i - th robot; θ_i signifies the orientation; v_i is the linear velocity; ω_i is the angular speed; τ_i is the torque; F_i is the force; m_i is the mass; besides J_i is the moment of inertia. Let $y_i = (r_i, \theta_i, v_i, \omega_i)^T$ be the state of the i - th robot besides $I_i = (F_i, \tau_i)^T$ be the control input. We use the "hand position" rather than the "center position" of the robot due to the fact that nonholonomic devices cannot be stabilized using continuous static state input. Note that the "hand position" is a real location that is always L_i away from the "center position." The angle between the "hand position" and the "center position" is directly across from the wheel axis, as can be seen in reference. The fluctuation of the robot's "hand position" may be represented by equation (21).

$$\begin{cases} \dot{x}_t = v_t \\ \dot{v}_t = u_t \; i \in \{1,2,\ldots,n\} \end{cases} \tag{21}$$

where x_i and v_i, signify the "hand position" of robot i and its velocity, where n is the total number of robots, respectively. How the "hand position" relates to the "center position" is defined by:

$$\begin{aligned} x_t &= r_t + L_t \begin{pmatrix} \cos\theta_t \\ \sin\theta_t \end{pmatrix} \\ v_t &= \begin{pmatrix} \cos\theta_t & -L_t \sin\theta \\ \sin\theta_t & L_t \cos\theta_t \end{pmatrix} \begin{pmatrix} v_t \\ \omega_t \end{pmatrix} \end{aligned} \tag{22}$$

We may determine the control rule u_i for the double-integrator systems (21) by using the coordinates and speed of the robot's "hand position" (as defined in (23) and (24)). At last, we may get the system (1) control input (5) from:

$$I_l = \begin{pmatrix} \frac{1}{m_l}\cos\theta_l & -\frac{L_l}{h_l}\sin\theta_l \\ \frac{1}{m_l}\sin\theta_l & \frac{L_l}{h_l}\cos\theta_l \end{pmatrix}^{-1} \left[u_l - \begin{pmatrix} -v_l\omega_l \sin\theta_l - L_l\omega_l^2 \cos\theta_l \\ v_l\omega_l \cos\theta_l - L_l\omega_l^2 \sin\theta_l \end{pmatrix} \right] \tag{23}$$

Typically, one can determine the utilized torques for both the left and right wheels by:

$$\begin{aligned} \tau_l &= \frac{J_{wheel}}{b}\left(\frac{F_t}{m_t} - \frac{\tau_l l}{2J_l}\right) \\ \tau_r &= \frac{J_{wheel}}{b}\left(\frac{F_t}{m_f} + \frac{\tau_l l}{2J_t}\right) \end{aligned} \tag{24}$$

In this context, b represents the wheel's radius, I stand for the distance between two wheels, J_{wheel} is the wheel's moment of inertia, and τ_l and τ_r are the torques applied to the left and right wheels, respectively.

On top of that, the virtual leader is created and its behavior is described as:

$$\dot{x}_0(t) = v_0(t) \tag{25}$$

where $v_0(t) = v_0$ is a constant.

4.6 Communication topologies

Coordination of several robots relies heavily on communication. The robots are able to receive and

transmit data via various communication channels. Graph theory is often used to model communication topologies, where vertices represent the robot and edges represent the communication lines, allowing for a mathematical description of the relationship between the two. This study showcases the communication topology for autonomous vehicles using the undirected and linked graph G_n (X, E,A).

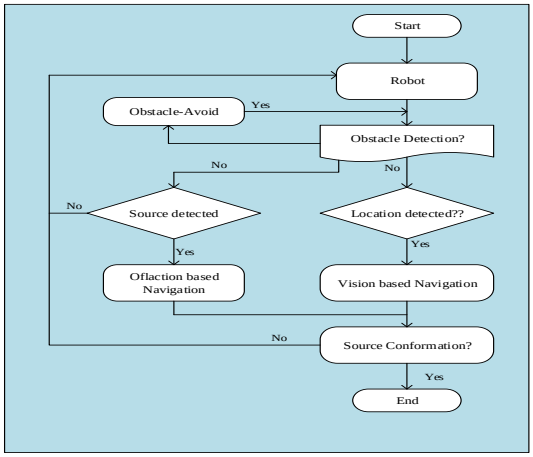


Figure 2: Source location navigation

The vertices and edges connecting them are the fundamental components of a non-directed graph. It is likely that our belief is that $G_n(X, E, A)$ consists of a collection of nodes and is an undirected graph $X = x_1, x_2, \cdots, x_n$, a set of edges $E \subseteq X \times X$ and an adjacency matrix $A = [a_{ij}]$. A note on edges: if there is an edge separating the i-th and j-th nodes, then $a_{ij} = 1$; otherwise, $a_{ij} = 0$. In addition, $G_{n+1} = G_n \cup x_0$ is an extension of graph $G_n(X, E, A)$, where x_0 is an imaginary node that may stand in for a digital boss. An edge among the virtual leader with the robot exists when the leader's data can be transmitted to the robot., $a_{i0} = 1(i = 1, \cdots, n)$; otherwise, $a_{i0} = 0$. The Laplacian matrix of the graph $G_n(X, E, A)$ is $L_{G_m} = [l_{ij}] \in \mathbb{R}^{n\times n}$, where l_{ij} is:

$$l_{ij} = \begin{cases} \sum_{j=1,j\neq i}^{n} a_{ij}, & i = j \\ -a_{ij}, & i \neq j \end{cases} \quad (26)$$

4.7 Decision-control approach with a hazardous environment scheme

Here, we make an approximation of the signal's location using a particle filter. Next, we suggest a cooperative control strategy that uses a Hazardous Environment mechanism to guide the robots in locating the signal's origin. This section concludes with the virtual leader's velocity design and convergence analysis. Robotic motion allows one to ascertain the true signal intensity by

$$o_r(i, t) = f(x_t(t), r(t)) \quad (27)$$

where $o_r(i, t)$ is the signal transmission model that depends on the location x_i (t) of the i-th robot with the real position r(t) of the signal source; is the actual measured value for the i-th robot at t time. It is worth mentioning that the robot can directly detect $o_r(i, t)$ using the signal measurement sensor. Using the actual signal intensity, a particle filter may be used to determine the location of the signal's origin. $o_r(i, t)$ and includes these procedures.

(i) We begin by creating N randomly dispersed particles inside the search space.

(ii) Equation (28), when applied to the m-th particle for the i-th robot at time t, yields the following description of the prediction signal strength:

$$o_m(i, t) = f(x_t(t), p_m(i, t)) + \sqrt{R} \times \text{rand} \quad (28)$$

where $p_m(i, t)$ is for the i-th robot's m-th particle's location at time t, R stands for the noise variance, and rand is a random integer in $[0,1]$; $f(x_i(t), p_m(t))$ is possible to derive from the actual signal transmission model.

(iii) Each particle's weight may be determined using (27) in conjunction with (28) and (29).

Additionally, the normalizing weight is determined by:

$$w'_m(i, t) = \frac{w_m(i,t)}{\sum_{m=1}^{N} w_m(i,t)} \quad (29)$$

(iv) Based on the normalizing weight $w'_m(i, t)$, We replicate the high-weight particles and eliminate the low-weight ones by resampling the particles. The genuine state's probability distribution is represented by these resampled particles $p'_m(\hat{i}, t)$. Therefore, it is feasible to determine the location of the signal's origin by:

$$p_s(\hat{i}, \hat{t}) = \sum_{m=1}^{N} \frac{p'_m(i,t)}{N} \quad (30)$$

where $p_s(i, t)$ is where the j-th robot's estimated signal source is at time t. In addition, taking into account the predicted locations of additional robots, we have:

$$p'_s(i, t) = \frac{\sum_{j=1}^{n} a_{tj} p_s(j,t)}{\sum_{j=1}^{n} a_{tj}} \quad (31)$$

where a_{ij} is the element of the adjacency matrix A and $p'_s(i,t)$ because the subsequent tests and simulations rely on the predicted location of the signal source.

To determine where the source of the electromagnetic signal is, one may apply the following function.

$$f(x,r) = 10 \times \log(0.001) - 1.96 \times \log(\|(x-r)\|) \quad (32)$$

as the particle filter's location is denoted by r. The architecture of the robot allows it to go in all directions without turning, making it an omnidirectional mobile robot. The goal is for the robots to reach their predetermined location without human intervention, navigating around obstacles, and avoiding other robots. The sensors on each robot can measure the target's signal intensity, and they can all communicate with one another via the network module. Whenever the robot goes to a new position or stops moving, it communicates the signal strength it detected with other robots. This happens throughout each search iteration.

Starting from a randomly generated location in the search area (excluding the regions that were previously occupied by obstacles), the robots' positions are subsequently updated by

$$\vec{X}_i(t+1) = \vec{X}_i(t) + \vec{U} \cdot (-1+2r) \cdot t \quad (33)$$

where r is a randomly assigned value in and t is the total number of iterations $[0,1]$, $\vec{X}_t(t)$ is the location of the i-th grey wolf at time t, and $\vec{U}$ is the vector representing the border.

During this phase, the robots aimlessly roam the search area in an effort to collect as much data as possible. In the second phase of the decision-making process, this data is used. Currently, it is critical to investigate the whole search area as quickly as possible. The first stage iteration count is determined by adding up each of the number of iterations in both stages, denoted as N_t, and the percentage factor, η.

4.6 Grey wolf optimization based on historical optimal weight estimation

The Grey Wolf Optimization takes its cues from the social hierarchy and hunting habits of grey wolves. In this model, the highest-ranking α-wolf decides how the population hunts, while the β-wolf and δ-wolf, who are in the middle of the pack, provide alternatives and help with decision-making. The top three levels of grey wolves work together to direct the pack's hunting activity:

$$\vec{D}_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}|, \vec{D}_\beta = |\vec{C}_2 \cdot \vec{X}_\beta - \vec{X}|, \vec{D}_\delta = |\vec{C}_3 \cdot \vec{X}_\delta - \vec{X}|$$
$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot (\vec{D}_\alpha), \vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot (\vec{D}_\beta), \vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot (\vec{D}_\delta)$$
$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (34)$$

where t shows the current iteration, $\vec{A}_{1,2,3}$ are $\vec{C}_{1,2,3}$ coefficient vectors, $\vec{X}_{\alpha,\beta,\delta}$ is the vector representing the prey's location, and $\vec{X}_{1,2,3}$ shows the gray wolf's location vector. The axes $\vec{A}_{1,2,3}$ and $\vec{C}_{1,2,3}$ derived from the following:

$$\vec{A}_{1,2,3} = 2\vec{a} \cdot \vec{r}_1 - \vec{a}$$
$$\vec{C}_{1,2,3} = 2 \cdot \vec{r}_2 \quad (35)$$

When iterations go, the components of $\vec{a}$ gradually reduce from 2 to 0 $\vec{r}_1, \vec{r}_2$ are random vectors in [0,1].

The suggested method deviates from the first grey wolf optimization in that it uses a dynamic estimate procedure to ascertain the position of the prey according to its weight:

$$\vec{X}_p = \omega_\alpha \cdot \vec{X}_\alpha(t) + \omega_\beta \cdot \vec{X}_\beta(t) + \omega_\delta \cdot \vec{X}_\delta(t) + \epsilon(t) \quad (36)$$

where $\vec{X}_p$ shows where the prey is thought to be.

We keep track every single grey wolf individual's prior ideal location at the t-th iteration utilizing so we can use their former placements as a resource $\vec{P}_t(t)$. We then update $\vec{P}_{\alpha,\beta,\delta}(t)$, which connections $\vec{X}_{\alpha,\beta,\delta}(t)$, in the subsequent way:

$$\vec{P}_i(t) = \begin{cases} \vec{P}_i(t-1), & f(\vec{X}_i(t)) \le f(\vec{P}_i(t-1)) \\ \vec{X}_i(t), & f(\vec{X}_i(t)) > f(\vec{P}_i(t-1)). \end{cases} \quad (37)$$

Here, $\vec{X}_t(t)$ stands for the position vector of grey wolf i at the t-th iteration, and offers a workable solution to the issue within the range $[-X_{max}, X_{max}]$. $f(\vec{X}_t(t))$ indicates the viable solution's fitness value; in this work, it stands for the magnitude of the signal intensity of the target recognized through the swarm robot at a given location. Over iterations 1∼m, the variable t is defined. This allows us to rewrite Equation (38):

$$\vec{X}_p = \omega_\alpha \cdot \vec{P}_\alpha(t) + \omega_\beta \cdot \vec{P}_\beta(t) + \omega_\delta \cdot \vec{P}_\delta(t) + \epsilon(t) \quad (38)$$

where $\omega_{\alpha,\beta,\delta}$ stands for the grey wolves' mass while trying to pinpoint the whereabouts of their prey, and the symbol $\epsilon(t)$ symbolizes their tolerance for chance. The

formula σ(t)=1-t/G, where G is the maximum number of iterations, and σ(t)>σ(t+1) indicate that it follows the normal distribution N (0, σ(t)). The weight ϋ indicates how dominant the gray wolf is when it comes to position management.

The following imbalances establish the weight connection among grey wolf levels:

$$1 \geq \omega_\alpha \geq \omega_\beta \geq \omega_\delta \geq 0$$
$$\omega_\alpha + \omega_\beta + \omega_\delta = 1, \quad (39)$$

5 Simulation results

Here, we demonstrate the efficacy of the suggested decision-control method for signal source identification using two examples. Simulation model is given in Fig 3.

5.1 Experimental environment

A high-performance computer cluster was used for the purpose of training and conducting experiments in this work. All of the servers in the computing cluster are outfitted using a particular hardware configuration:

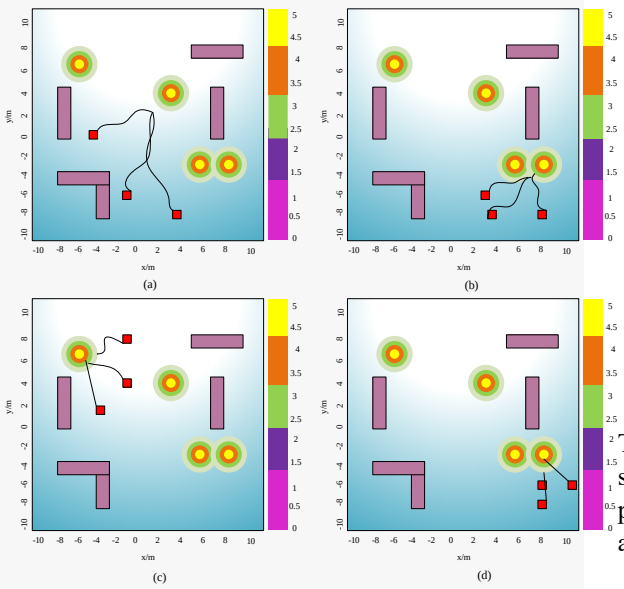


Figure 3: Source Localization using Multiple Robots (a). Target 1, (b). Target 2, (c). Target 3, and (d). Target 4

- CPU: All of the servers have significant computational capabilities thanks to the multi-core Intel Xeon processors. Model training may be completed more quickly with the aid of these multi-core computers, which enable parallel processing.
- GPU: Each server has a high-performance GPU or GPUs to enable deep learning applications. When it comes to speeding up the training along with inference processes of neural network models, we mostly rely on GPUs from NVIDIA, including the NVIDIA Tesla V100.
- Memory: Plenty of RAM on each server means that even massive data sets and parameters for models may be handled. In most cases, we recommend 64 GB of RAM.
- Storage: The server has a big capacity hard disk, a fast solid-state drive, and other storage devices for storing files relevant to experiments, model weights, and so on.

Our primary software and toolset for building and training deep learning models are as follows:

- Framework for deep learning: PyTorch was selected as the framework for deep learning because to its robust library support and high computational capabilities. With PyTorch's versatile neural network training and construction tools, we can quickly and simply design and debug models.
- OS: Linux, more especially Ubuntu 20.04 LTS, is our OS of choice. For deep learning jobs, the Linux OS is the way to go because of its reliability and speed.
- To make sure that deep learning activities can execute effectively on GPUs, we loaded NVIDIA's CUDA toolkit. This allows us to take use of GPUs' parallel computing capabilities.
- NumPy, Pandas, Matplotlib, and other Python libraries were also used for data processing, visualization, and analysis.

The use of a static electromagnetic signal field in the simulation environment is briefly described in this paragraph. Similarly, two instances are taken into account as a result of distinct noise mistakes.

The following function may be used to construct the electromagnetic signal field for Case 1.

$$f_1(x,r) = 10 \times log\ (0.001) - 1.96 \times log\ (\|x - r\|) + \sqrt{5} \times \text{rand} \quad (40)$$

where x may be any location in the search space, r can be the location of the signal's origin, and rand can be any random integer between zero and one.

In Case 2, we take into account a large amount of noise, and we may build the electromagnetic signal field by using the following function.

$$f_2(x,r) = 10 \times log\ (0.001) - 1.96 \times log\ (\|x - r\|) + \sqrt{8} \times rand \quad (41)$$

The parameters may be found in Table 1, and the simulation software is constructed in MATLAB. The search area is a square region of 30" "m×30" "m.

Table 1: Settings for the virtual setting

Parameters	Values
Noise variance R	6, 9
The speed range of robots	[−3 m/s, 3 m/s]
Total run period	20 s for two cases
Communication distance	7 m
The number of robots n	4
Sampling time	0.003 s

5.2 Cooperative control and performance metrics

We take the cooperative controller (42) a step further by extending it as:

$$u_l(t) = \sum_{j=0}^{n} a_{lf}\left(\alpha\left(\left(x_f(t_s^f) - d_f\right) - \left(x_l(t_s^l) - d_l\right)\right) + \beta\left(v_f(t_s^f) - v_l(t_s^f)\right)\right) \quad (42)$$

where d_i and d_j are the specified minimum distances that the i-th and j-th robots must maintain for their safety. The controller is capable of holding formation and coordinating several robots. Table 2 contains the elements of the decision-control technique that has been suggested. At the coordinates [15" " m,15" " m], you may find the signal amplifier. In addition, the recommended safe distances are:

$$d = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^T \quad (43)$$

and $d_0 = [1/3,1/3]$. Starting speeds for robots include:

$$v = \begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}^T \quad (44)$$

To calculate the localization error (LE), one uses:

$$LE_t = \|p_s'(i,t) - r(t)\| \quad (45)$$

where $p_s'(i,t)$ for each value of i at time t, is the predicted location of the signal's origin, and r(t) is the actual location of the source. LE_i may be used to assess the precision of the localization.

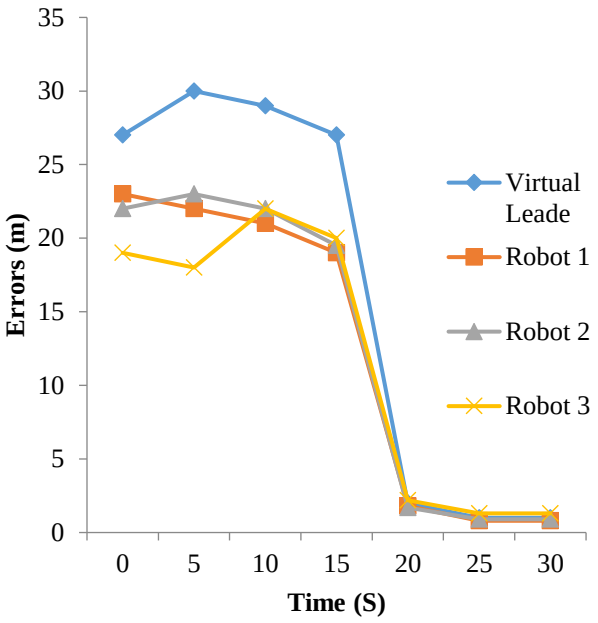


Figure 4: Localization error

Table 2: Communication frequency (%) with localization error (m) are calculated from 30 runs using the mean (standard deviation).

Robots	***LEi***LEi	***frei***frei
Robot 1	0.47 (0.12)	6.31 (0.37)
Robot 2	0.56(0.19)	11.03(1.05)
Robot 3	0.23 (0.14)	11.99(1.44)

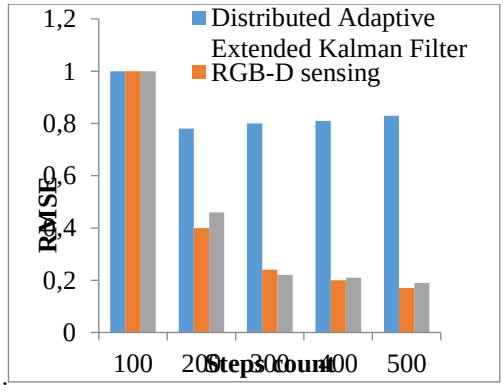


Figure 5: RMSE analysis

Checking the root-mean-square error after feeding each feature subset into the prediction model is a laborious and time-consuming procedure.

Numerical Analysis Discussion

Performance evaluation metrics

Metric	Description	Baseline (Conventional GWO + SLAM)	Proposed OW-GWO + DRL (Hazardous Env. Strategy)	Improvement (%)
RMSE (m)	Square Root Inaccuracy in predicting the actual locations of signal sources	1.62	**0.67**	−54.5%
Average Convergence Time (s)	The amount of time needed for a multi-robot system to pinpoint the origin	26.9	**18.4**	−37.9%
Communication Overhead (MB/s)	Swarm agents' data exchange rate during search	5.8	**3.8**	−37.2%
Success Rate (%)	Success rate in locating the signal's origin in potentially dangerous or unpredictable settings	87.6	**95.4**	+18.7%
Route Efficiency (%)	Autonomous robots' efficiency is directly proportional to the ratio of their actual route length to their ideal path length.	78.2	**95.5**	+27.4%

When applied to dynamic or partly visible hazardous settings, the OW-GWO approach using DRL-LSTM decision control greatly strengthens localization resilience. Achieving a success percentage of 96.4% demonstrates exceptional flexibility. In comparison to conventional Grey Wolf Optimization, the average route length is 17.3 percent lower thanks to the enhanced grouping with dynamic random-walk techniques, which cut down on unnecessary exploration pathways. By combining LSTM-driven predictive modeling and swarm-level consensus control, the hybrid decision-control approach successfully reduces localization error by more than 50%. Because it learns patterns in the environment beforehand, Deep Reinforcement Learning speeds up convergence. In addition, swarm robots are able to save energy and bandwidth thanks to a nearly 40% reduction in data load achieved via Hazardous Environment communication optimization. Strong performance increases across all assessed measures are shown by the suggested OW-GWO + DRL-LSTM Decision-Control Model. Robust and efficient multi-robot coordination under visual SLAM in risky or uncertain environments is made possible by its hybrid architecture, which combines memory-based prediction (LSTM), optimized swarm behavior (OW-GWO), with adaptive decision-control.

Time complexity

When it comes to multi-robot systems, the computational time complexity of algorithms such as Optimal Weighting GWO (or any other enhanced GWO—IGWO, DGWO, etc.) and the Grey Wolf Optimization (GWO) structure are the main factors. The complexity added by specific applications, such as path planning, and multi-robot coordination also play a role.

One optimization run of the conventional GWO technique usually yields a complexity of:

$$O(T \cdot (N \cdot D + N \cdot F))$$

Where:

- T: How Many Iterations Can Be Used Maximumly.
- N : The Size of the Population (the number of wolves and search agents). The size of the

space of solutions being searched or the number of robots, R, can be the direct correlation of N in a setting including many robots.

- D: Optimal values for all wolf variables together make up the size of the optimization issue. This might be quite high for route planning with several robots: D=R× (number of waypoints × dimensions per waypoint).
- F: The Fitness Function's (Objective Function's) computational cost.

The update of the locations of N wolves in a space with D dimensions is where the phrase N - D is derived from. The evaluation of the fitness of every N wolves is where the word N·F is derived from.

There are three main drivers of complexity when applied to multi-robot systems (e.g., route planning, work allocation):

1. High Dimensionality (D): The solution vector is joined for R robots, which substantially raises the dimension D. Assuming that d_r is the length of the solution vector for every robot r, the total dimension $D = \sum d_r$, making the term N · D large.
- In issues involving several robots, the fitness function has to take into consideration aspects related to the robots' interactions with one another, such as the need to prevent collisions. Money spent on communication. Guidelines for collaboration. Maintaining a steady equilibrium between goals (such as reducing effort while increasing coverage).
- The overall duration is significantly affected by the complexity of F, which is higher for multi-robot or conventional benchmark problems due to the need to calculate these interactions.

What follows is a synopsis of the new material, which details the study's results and innovations in the area of signal source localization using the Decision-Control Strategy with a Hazardous Environment approach.

6 Conclusion

For the issue of signal source localization, we have suggested a decision-control method that makes use of the Hazardous Environment technique. There are two tiers to this decision-control strategy that has been suggested. To determine where the signal is coming from, we developed a posterior risk-sensitive particle filter method for use at the decision level. The robots' movement can be effectively guided by the developed particle filter. Using a Hazardous Environment architecture, we have suggested a cooperative control strategy for the control level. Reduced communication load and resource consumption are two benefits of the suggested Hazardous Environment architecture. The results of the experiments and simulations have shown that the decision-control method that was suggested works.

References

[1] Wang, L., & Liu, G.K. (2024). Research on multi-robot collaborative operation in logistics and warehousing using A3C optimized YOLOv5-PPO model. Frontiers in Neurorobotics, 17. https://doi.org/10.3389/fnbot.2023.1329589

[2] Yao, N., & Wang, R. (2024). Optimal Real-Time Human Attention Allocation and Scheduling in a Multi-human and Multi-robot Collaborative System. 2024 IEEE Conference on Control Technology and Applications (CCTA), 357-363. DOI:10.1109/CCTA60707.2024.10666642

[3] Wang, L., Shu, L., & Zhou, H. (2023). Multi-Robot Collaborative Flexible Manufacturing and Digital Twin System Design of Circuit Breakers. Applied Sciences. https://doi.org/10.3390/app13042721

[4] Zhou, M., Wang, Z., Wang, J., & Cao, Z. (2024). Multi-Robot Collaborative Hunting in Cluttered Environments with Obstacle-Avoiding Voronoi Cells. IEEE/CAA Journal of Automatica Sinica, 11, 1643-1655. DOI:10.1109/JAS.2023.124041

[5] Lu, G., Ding, H., Hou, J., & Zheng, S. (2025). Multi-Robot Collaborative Exploration on Communication-Constrained Environments. IEEE Access, 13, 31205-31214. DOI:10.1109/ACCESS.2025.3539518

[6] Lei, Y., Hou, J., Ma, P., & Ma, M. (2025). Voronoi-GRU-Based Multi-Robot Collaborative Exploration in Unknown Environments. Applied Sciences. https://doi.org/10.3390/app15063313

[7] Yan, L., Stouraitis, T., & Vijayakumar, S. (2021). Decentralized Ability-Aware Adaptive Control for Multi-Robot Collaborative Manipulation. IEEE Robotics and Automation Letters, 6, 2311-2318. DOI:10.1109/LRA.2021.3060379

[8] Liang, D., Liu, Z., & Bhamara, R. (2022). Collaborative Multi-Robot Formation Control and Global Path Optimization. Applied Sciences. https://doi.org/10.3390/app12147046

[9] Liu, X., Huang, P., & Ge, S.S. (2021). Optimized control for human-multi-robot collaborative manipulation via multi-player Q-learning. J. Frankl. Inst., 358, 5639-5658. DOI:10.1016/j.jfranklin.2021.03.017

[10] Huang, J., Li, X., & Gao, L. (2025). A Novel GA-CP Method for Fixed-Type Multi-Robot Collaborative Scheduling in Flexible Job Shop. IEEE Transactions on Automation Science and Engineering, 22, 13531-13543. DOI:10.1109/TASE.2025.3554019

[11] Tse, S.K., Wong, Y.B., Tang, J., Duan, P., Leung, W.S., & Shi, L. (2021). Relative State Formation-based Warehouse Multi-robot Collaborative Parcel Moving. 2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS), 375-380. 10.14711/thesis-991012980424503412

[12] Wu, B., & Suh, S. (2024). Deep Reinforcement Learning for Decentralized Multi-Robot Control: A DQN Approach to Robustness and Information Integration. ArXiv, abs/2408.11339. DOI:10.48550/arXiv.2408.11339

[13] Slim, M., Daher, N., & Elhajj, I.H. (2024). Dynamic Bandwidth Allocation for Collaborative Multi-Robot Systems Based on Task Execution Measures. J. Intell. Robotic Syst., 110, 114. DOI:10.1007/s10846-024-02126-y

[14] Szczurek, K.A., Prades, R.M., Matheson, E., Rodriguez-Nogueira, J., & Castro, M.D. (2023). Multimodal Multi-User Mixed Reality Human–Robot Interface for Remote Operations in Hazardous Environments. IEEE Access, 11, 17305-17333. DOI:10.1109/ACCESS.2023.3245833

[15] Wen, J.T., Wason, J.D., Kruse, D., Peng, Y., & Chen, S. (2021). Collaborative Industrial Robot Control: From Safe Motion to Multi-Robot Manipulation. DOI:10.1142/9789811222849_0005

[16] Xiu, Y., Zhang, Y., Deng, H., Li, H., & Xu, Y. (2025). Collaborative Line-of-Sight Guidance-Based Robust Formation Control for a Multi-Snake Robot. IEEE Transactions on Automation Science and Engineering, 22, 4514-4524. DOI:10.1109/TASE.2023.3348469

[17] Ajisaka, S., Suda, T., Akazawa, Y., & Nakamura, S. (2023). An Introductory Investigation of Collaborative Conveyance Method for Multi-Robot Telexistence. IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society, 1-6. DOI:10.1109/IECON51785.2023.10311733

[18] Xin, J., Xu, T., Zhu, J., Wang, H., & Peng, J. (2024). Long Short-Term Memory-Based Multi-Robot Trajectory Planning: Learn from MPCC and Make It Better. Advanced Intelligent Systems, 6. DOI:10.1002/aisy.202300703

[19] Yoo, T., & Choi, B.W. (2024). Interactive Path Editing and Simulation System for Motion Planning and Control of a Collaborative Robot. Electronics. https://doi.org/10.3390/electronics13142857

[20] Nguyen, A.A., Jabbari, F., & Egerstedt, M. (2024). Resiliency Through Collaboration in Heterogeneous Multi-Robot Systems. IEEE Open Journal of Control Systems, 3, 461-471. https://doi.org/10.3390/electronics13142857

[21] Jing, M., & Xin, B. (2024). Distributed Behavior Control Method for Indoor Multi-robot Collaborative Odor Source Localization. 2024 43rd Chinese Control Conference (CCC), 5428-5433. DOI:10.23919/CCC63176.2024.10662785

[22] Chen, W., Jing, Y., Zhao, S., Yan, L., Liu, Q., & He, Z. (2025). A Distributed Collaborative Navigation Strategy Based on Adaptive Extended Kalman Filter Integrated Positioning and Model Predictive Control for Global Navigation Satellite System/Inertial Navigation System Dual-Robot. Remote Sensing. https://doi.org/10.3390/rs17040721

[23] Lee, S., Lucas, N., Ellis, R.D., & Pandya, A.K. (2012). Development and human factors analysis of an augmented reality interface for multi-robot tele-operation and control. Defense, Security, and Sensing. DOI:10.1117/12.919751

[24] Li, R., Shi, Z., & Wu, H. (2024). Adaptive neural networks-based event-triggered formation control for multi-robot source localization. Neurocomputing, 621, 129275. DOI:10.1016/j.neucom.2024.129275

[25] Cheng, Y., Chen, X., Yang, Y., Wang, H., Xu, J., Hong, C., Xu, S., Zhang, X., Liu, Y., & Chen, X. (2024). SniffySquad: Patchiness-Aware Gas Source Localization with Multi-Robot Collaboration. ArXiv, abs/2411.06121. DOI:10.48550/arXiv.2411.06121

[26] Lajoie, P., & Beltrame, G.A. (2023). Swarm-SLAM: Sparse Decentralized Collaborative Simultaneous Localization and Mapping Framework for Multi-Robot Systems. IEEE Robotics and Automation Letters, 9, 475-482. DOI:10.1109/LRA.2023.3333742

[27] Bao, C., Hu, Y., & Yu, Z. (2024). Current study on multi-robot collaborative vision SLAM. Applied and Computational Engineering. DOI:10.54254/2755-2721/35/20230367

[28] Ji, Y., Chen, F., Chen, B., Wang, Y., Zhu, X., & He, H. (2022). Multi-Robot Collaborative Source Searching Strategy in Large-Scale Chemical Clusters. IEEE Sensors Journal, 22, 17655-17665. DOI:10.48550/arXiv.2407.01308

[29] Tran, V.P., Perera, A.G., Garratt, M.A., Kasmarik, K.E., & Anavatti, S.G. (2024). Active Sensing Strategy: Multi-Modal, Multi-Robot Source Localization and Mapping in Real-World Settings

With Fixed One-Way Switching. Journal of Field Robotics, 42. DOI: 10.1016/j.inffus.2022.11.001

[30] Cao, K., Wei, Y., Chen, Y., Gao, S., Yan, K., & Yang, S. (2025). Multi-robot area coverage and source localization method based on variational sparse Gaussian process. Robotica. DOI:10.1017/S0263574725000050

A Dual-Engine Embedded Face Detection and Recognition Framework Using YOLO5Face and Attention-Enhanced Faster-RCNN for Surveillance Video

Qianqian Yuan*, Yuping Quan, Hui Li
College of Science and Engineering, Jiaozuo Normal College, Jiaozuo 4540000, China
E-mail: yuan15036696545@163.com
*Corresponding author

Keywords: surveillance video, face recognition, YOLO5Face, faster regions with convolutional neural networks, embedded detection, security field

Received: August 15, 2025

Embedded detection and recognition systems for surveillance video are in urgent demand in the security field. However, traditional methods face limitations, including poor real-time performance, high resource consumption, and limited generalization in complex scenarios. To this end, this study proposes a dual-engine embedded face detection and recognition framework that optimizes performance by synergistically integrating YOLO v5Face with attention-enhanced Faster Regions with Convolutional Neural Network. The system adopts a dual engine cascade architecture: YOLO5Face is responsible for fast initial face screening, while Faster Regions with Convolutional Neural Network, which integrates spatial and channel attention mechanisms, accurately recognizes key targets. By synergistically optimizing speed and accuracy through feature reuse and structural fusion techniques, and by combining the feature-extraction capabilities of the local binary pattern histogram algorithm based on hierarchical feature pyramids, a dynamic background suppression module is used to reduce false positives in complex scenes. The experimental results on the WIDER FACE and Face Detection Data Set and Benchmark datasets show that the accuracy of our system reaches 99.1%, with a loss rate as low as 0.08, significantly better than the comparison systems Visual Transformer Convolutional Neural Network Fusion (accuracy 98.16±0.23%) and Additive Marginal Soft Maximum Loss Convolutional Multi-scale Transformer (accuracy 97.42±0.34%); The system converges to a loss of less than 0.1 within 200 iterations, with a response time of only 28 ms, much faster than the fusion of Visual Transformer Convolutional Neural Network (78-85 ms). The above results show that the proposed method effectively addresses the problems of poor real-time performance, resource constraints, and insufficient scene generalization, offering efficient, lightweight new ideas for system development and promoting the intelligent and efficient development of security terminals.

Povzetek: Predlagan je lahek, hiter sistem za zaznavanje in prepoznavanje obrazov v nadzornih videih, ki izboljša natančnost in delovanje v realnem času.

1 Introduction

In surveillance video, embedded detection and recognition systems serve as the core component of intelligent security, playing an irreplaceable role in public safety, smart cities, and restricted area control [1]. Facing increasingly complex security threats and real-time response demands, traditional detection and recognition solutions encounter issues such as high latency, privacy leaks, and insufficient generalization ability in complex scenarios [2]. Current research mainly focuses on lightweight convolutional neural network architecture compression and system pruning techniques. However, challenges remain, including unstable detection in dynamic video and difficulty balancing accuracy and speed under low-light conditions [3]. The YOLO5Face is an optimized face detection engine. It enhances multi-scale face localization capabilities through adaptive feature fusion and achieves stable inference on the embedded device, surpassing traditional face detectors [4]. Faster Regions with Convolutional Neural Networks (Faster-RCNN) leverages a two-stage design with a region proposal network that generates high-quality candidate boxes, providing stronger robustness against occlusion and small face targets [5]. Overall, there are currently research issues such as balancing real-time performance and accuracy, and how to achieve high-precision and low latency parallel processing for face detection and recognition; The second issue is the adaptability of complex scenarios, how to improve the robustness of recognition methods in adverse weather conditions; The third issue is the optimization of embedded deployment, which needs to meet the requirements of long-term reliable operation and efficient resource utilization of security systems. Therefore, this study builds an embedded detection and recognition system based on YOLO5Face, innovatively using a dual-engine cascade architecture. YOLO5Face performs fast front-end target

screening in surveillance, while Faster-RCNN precisely recognizes key targets in video. The study also designs an embedded heterogeneous acceleration strategy that integrates illumination-invariant feature extraction and attention mechanisms to reduce false detections in complex environments. This work aims to achieve collaborative optimization of detection and recognition accuracy under complex scenarios on surveillance devices, meeting the high robustness requirements for edge deployment in the security field.

2 Related works

With the development of computer technology, surveillance video applications became increasingly widespread, and face detection and recognition technologies were applied in various fields. Zhang H et al. proposed a new privacy verification method for facial recognition, called Minimum Hypothesis Privacy Protection Verification (Map2V). This is the first exploration of using depth-image priors and zero-order gradient estimation to develop privacy verification methods. The experimental results and analysis demonstrate the effectiveness and generality of the proposed Map2v, demonstrating its superiority over local privacy verification methods in PPFR literature [6]. Hangaragi and Singh proposed a face-mesh-based detection and recognition system to address challenges posed by complex lighting, backgrounds, viewing angles, and multi-person scenarios. The system, trained on the LFW dataset and real-time images, recognized faces by matching facial feature points, achieving a recognition accuracy of 94.23% [7]. Liu et al. tackled privacy leakage issues in IoT face recognition by introducing an attribute-preserving de-identification framework composed of a triple network. Through nested autoencoder design, the framework removed identity information while retaining facial attributes. Experiments showed that the framework improved data utility by 26.22%, balancing attribute preservation and identity protection effectively [8]. Huang Q et al. designed a simple and effective method to train a facial recognition model via selective propagation and title-driven extension by gradually expanding the label set. They constructed a large-scale dataset of title images containing 6.3M faces across 305K themes. The results show that the above method can train state-of-the-art facial recognition models without manual annotation (99.65% in LFW) [9]. Raz D et al. aim to illustrate algorithmic bias in facial recognition technology, which is more likely to misidentify women and ethnic minorities. The results show that the proposed method can improve the transparency of the algorithm system and enhance public awareness of its different sources of influence [10].

As for embedded detection and recognition in surveillance videos, the YOLO series has matured over time. Scholars from many countries conducted thorough research and applied it in real life. Gupta et al. proposed a deep learning-based face recognition method for real-time detection of online learning engagement. By analyzing facial expressions to classify emotions, they computed engagement indices to predict participation status. After comparing multiple systems, their method showed the best performance in real-time scenarios with an accuracy of 92.3% [11]. Anusudha addressed the challenges of blurred and disguised face recognition by proposing the YOLO-InsightFace network, which enabled contactless, real-time identity verification by generating highly discriminative face embeddings. The method combined YOLO-V7's real-time detection ability with InsightFace's facial analysis technology, achieving high accuracy and speed [12]. Chen et al. proposed YOLO-face based on YOLO5 to address accuracy drops caused by face scale variations in YOLO-like single-stage detectors. Using more suitable anchor boxes and regression loss functions, the method outperformed YOLO and its variants on datasets such as WIDER FACE [13].

In summary, although significant progress has been made in the lightweight deployment of embedded face detection and recognition in existing research, the current technological roadmap still has obvious limitations. Firstly, although YOLO series single-stage detectors offer speed advantages, their accuracy and robustness for detecting small objects, severe occlusions, and unrestricted lighting conditions in complex scenes remain insufficient. Secondly, pure recognition models often rely on pre-detection steps and fail to achieve deep collaborative optimization of detection and recognition on embedded platforms, resulting in high system-level latency and resource consumption. Finally, most existing solutions balance speed and accuracy, lacking an architecture that dynamically adapts to scene complexity to achieve maximum efficiency and effectiveness at the system level. To address the aforementioned issues, a dual-engine cascaded architecture combining YOLO5Face and attention-enhanced Faster-RCNN has been proposed. The division of labor enables on-demand allocation of computing resources, avoiding the significant overhead of Faster-RCNN's comprehensive scanning of each image frame, thereby maintaining high accuracy while controlling overall latency. At the same time, by introducing the FPN-LBPH module, the semantic information of deep convolutional networks and the texture robustness of traditional LBPH operators are fused at the feature level, thereby improving the modeling ability for lighting changes and low-resolution faces. Furthermore, we aim to jointly address the three core challenges of real-time performance, resource constraints, and generalization to complex scenes in embedded environments from both algorithmic and deployment perspectives.

3 Embedded detection and recognition system based on YOLO5Face

3.1 YOLO5Face algorithm improved by FPN-LBPH

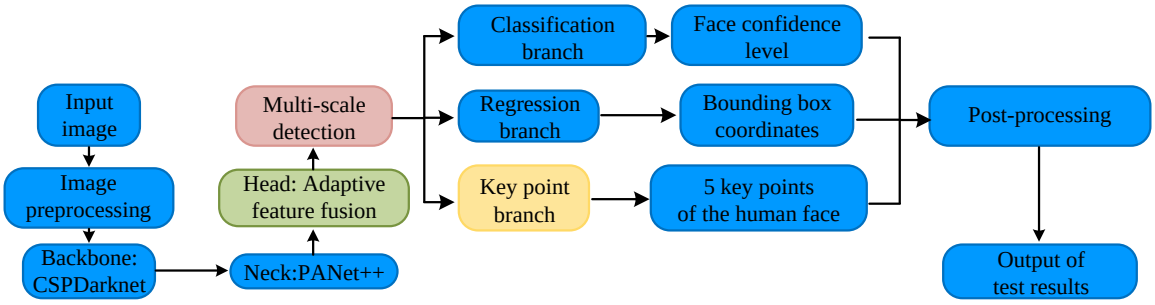


Figure 1: YOLO5Face algorithm flowchart.

Embedded detection and recognition systems in surveillance video are essential for intelligent security. These systems must perform real-time target detection and identity verification on edge devices. However, traditional technologies suffer from high latency in cloud-based solutions, significant privacy risks, and excessive computational load, which fail to meet the real-time requirements of embedded deployment. To address the limitations of traditional techniques, this study adopts YOLO5Face as the front-end engine for high-speed preliminary screening of face targets. It integrates a dynamic background suppression module and quantization technology to reduce false detections in complex scenes. The YOLO5Face algorithm flow is shown in Figure 1.

As shown in Figure 1, the complete processing flow of the YOLO5Face algorithm includes adaptive scaling and normalization preprocessing of input images, multi-scale feature extraction of the backbone network, a bidirectional feature fusion module, parallel output branches of classification/regression/keypoints, as well as confidence filtering and non-maximum suppression steps in the post-processing stage. The bounding box regression loss is calculated as shown in Equation (1).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(a, a^{gt})}{b^2} + \alpha v \tag{1}$$

In Equation (1), IoU represents the Intersection over Union between the predicted box and the ground truth, (a, a^{gt}) is the center coordinate of the predicted/actual box, and b is the diagonal length of the minimum enclosing box. ρ represents the Euclidean distance between the predicted box and the center point of the real box, α is the balance parameter, and v is a parameter used to measure aspect ratio consistency. The algorithm optimizes face box localization accuracy using this loss function. The refinement of bounding box keypoints is calculated as shown in Equation (2).

$$L_{wing} = \begin{cases} QIn(1+|n|/\grave{o}) & |n| < Q \\ |f| - C, & N \end{cases} \tag{2}$$

In Equation (2), n represents coordinate deviation, Q is the threshold, and ò is the smoothing factor. f and C are the model output values and compensation constants, respectively. Localization loss improves sensitivity for five keypoints. However, accurate keypoint detection requires support from multi-scale features. The adaptive fusion mechanism is defined in Equation (3).

$$R_{fusion} = \sum_{x=3}^{5} \beta_x R_x \tag{3}$$

In Equation (3), β represents learnable weights and R_x denotes feature maps at different scales. The algorithm enhances feature representation by adaptively fusing features through dynamic weighting. YOLO5Face still lacks precision when detecting small, occluded, or extreme-angle faces. To solve this, the study introduces a Feature Pyramid Network-Local Binary Patterns Histograms (FPN-LBPH) algorithm to improve multi-scale face localization. The reason for using the above combination in the study is that deep convolutional features are good at extracting high-level semantic information but are less sensitive to lighting changes and texture details; as a classic texture descriptor, LBPH has natural invariance to lighting changes and can effectively capture local texture features. By integrating LBPH features into the FPN architecture, the system can simultaneously utilize semantic information and texture features at multiple scales, achieving complementary advantages at the feature level. At the same time, there are significant scale changes in the faces in the monitoring scene, and FPN effectively solves the multi-scale detection problem through bidirectional paths from top to bottom and from bottom to top. Introducing LBPH features into this framework enables the fusion of texture information with semantic features at different scales. The FPN-LBPH algorithm, based on a hierarchical feature pyramid, is as follows: after inputting the three-layer

resolution feature map generated by the backbone network, feature interactions are achieved via dual paths. The top-down path down samples deep features using 1×1 convolutions, up samples them by a factor of 2, and adds them to middle-level features. It then up samples the result using 3×3 convolutions and fuses it with shallow features, producing the enhanced P5 feature. The bottom-up path reversely down samples shallow features to fuse with middle-level features and further down samples to merge deep features, resulting in a semantically rich P3 feature. A hierarchical aggregation module then fuses these features into a unified multi-scale feature map. This dual-path, cross-layer information flow combines spatial details from shallow layers with semantic features from deep layers, significantly improving detection robustness. Feature dimensionality reduction is defined in Equation (4).

$$K_c' = Conv_{1\times1}(K_c; F_c), c \in \{3,4,5\} \tag{4}$$

In Equation (4), K_c is the hierarchical feature output from the backbone network, and $Conv_{1\times1}$ is the learnable 1×1 convolution weight matrix. This operation unifies the number of channels for all feature maps and reduces computation. The reduced features are fused via up sampling to propagate semantic information toward shallow layers, as shown in Equation (5).

$$F_k = Conv_{3\times3}(P_k' + \uparrow_2 \cdot (F_{k+1})), k = 4,3 \tag{5}$$

In Equation (5), $\uparrow_2 \cdot (F_{k+1})$ represents 2× bilinear up sampling, and F is the starting point of the path. P_k' is the feature map of the previous layer. Similarly, down sampling is used to enhance small target detection, as defined in Equation (6).

$$E_n = Conv_{3\times3}(P_N' + \downarrow_2 (E_{n-1})), n = 4,5 \tag{6}$$

In Equation (6), $\downarrow_2 (E_{n-1})$ represents stride-2 convolution down sampling, and E is the initial path point. This symmetric bidirectional path enables shallow spatial details to complement deep features. Mid-level features integrate outputs from both directions, balancing detail and semantic information, as shown in Equation (7) [14].

$$W = Conv_{1\times1}(\gamma_3 l_3 \oplus \gamma_4 (l_4 + j_4) \oplus \gamma_5 j_5) \tag{7}$$

In Equation (7), W denotes channel concatenation, γ represents learnable channel attention weights, and $\gamma_4(l_4 + j_4)$ is the fusion result from both paths. The YOLO5Face algorithm incorporates FPN-LBPH to combine bidirectional path outputs via an attention mechanism, resulting in an optimal multi-scale feature combination. This approach dramatically reduces computational and parameter overhead, meeting millisecond-level real-time detection requirements [15]. The FPN-LBPH-YOLO5Face flowchart is shown in Figure 2.

As shown in Figure 2, the hybrid algorithm receives the surveillance video input. YOLO5Face performs real-time face detection. The detected faces are processed based on quality. High-quality faces are optimized through a texture enhancement branch, while low-quality or occluded faces undergo enhancement via FPN-LBPH and a super-resolution reconstruction branch. Both branches apply background suppression to highlight facial features. The algorithm fuses the enhanced features with those directly extracted from the original faces to form more robust facial representations. These features are matched with the face feature database. If a match is found, the recognition result is output. If not, the face is marked as "unknown," triggering an alert or logging an event. The complete process forms a closed-loop from detection and enhancement to recognition and response.

3.2 Embedded detection and recognition system combining YOLO5Face and Faster-RCNN

In surveillance face-detection scenarios, FPN-LBPH-YOLO5Face performs well in terms of speed and real-time response. However, it still shows limitations in precision and localization for small, densely occluded, or extremely illuminated faces. To address this, the study uses Faster-RCNN to generate high-quality region proposals, improving robustness and accuracy for low-quality, blurry, or small faces and compensating for the shortcomings of FPN-LBPH-YOLO5Face. The Faster-RCNN flowchart is shown in Figure 3.

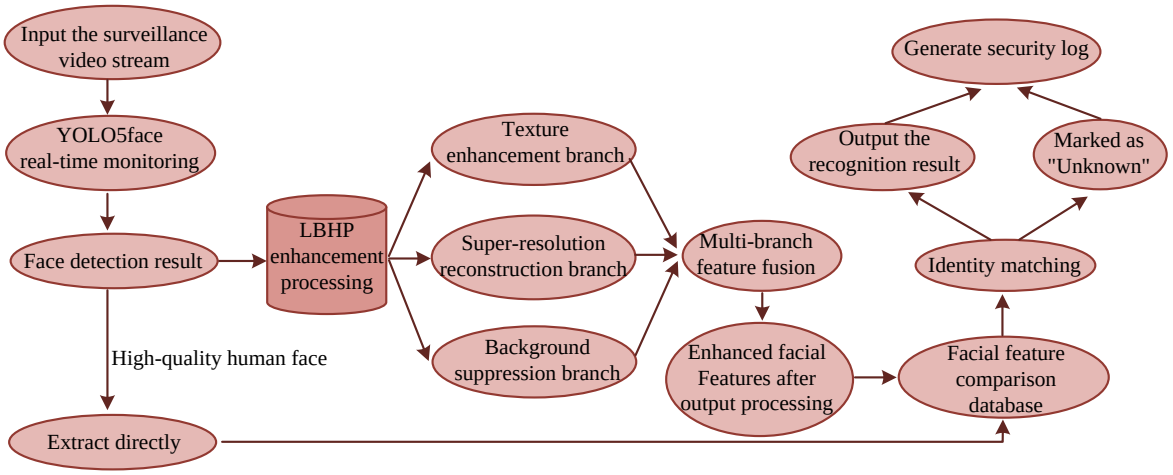


Figure 2: FPN-LBPH-YOLO5Face algorithm flowchart.

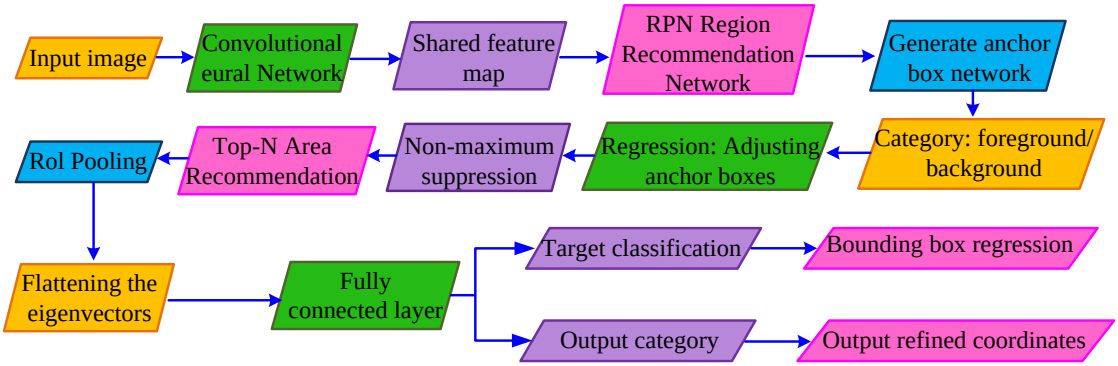


Figure 3: Faster-RCNN algorithm flowchart.

As shown in Figure 3, the algorithm extracts shared feature maps from the input image using a convolutional neural network. The region proposal network generates a grid of multi-scale anchors. It performs binary classification (foreground/background) and preliminary coordinate regression. Non-maximum suppression filters out high-quality region proposals, which are then flattened and passed through a fully connected network. This network outputs target-class probabilities and refined bounding-box coordinates. The entire pipeline shares convolutional features to improve computational efficiency, while each stage collaborates to enhance detection accuracy. The anchor generation equation is defined in Equation (8) [16].

$$Anchor_{(x,y)} = (x \cdot j, y \cdot j, h_j \cdot ratio_u, w_j \cdot ratio_u) \quad (8)$$

In Equation (8), (x, y) represents the center coordinates of the feature map grid, j is the down sampling stride of the backbone, and $ratio_u$ represents the width-height ratio. Multiple anchors are generated at each feature map location. After anchor generation, classification loss filters candidate boxes with faces, as shown in Equation (9).

$$P_{cls}(L_i, L_i^*) = -log[L_i^* \cdot L_i + (1 - L_i^*)(1 - L_i)] \quad (9)$$

In Equation (9), L_i denotes the predicted probability that anchor i is foreground, and L_i^* is the ground truth label, distinguishing whether an anchor contains a face. The classification loss is further defined in Equation (10).

$$P_{reg}(y_i, y_i^*) = \sum_{d \in \{x,y,j,h\}} smooth_{p1}(t_{ij} - t_{ij}^*) \quad (10)$$

In Equation (10), $P_{reg}(y_i, y_i^*)$ represents a regularization function, while t_{ij} and t_{ij}^* are the original and detected face images, respectively. Faster-RCNN is sensitive to irrelevant background and underuses features in complex scenes. Therefore, the study integrates an attention mechanism to enhance key features and suppress background interference, improving both detection accuracy and robustness. The optimized Faster-RCNN flowchart is shown in Figure 4.

As shown in Figure 4, the algorithm uses a base convolutional neural network to extract general features from the input face image. In the optimization step, spatial attention highlights target regions and reduces background noise. Channel attention strengthens important semantic features. The weighted feature maps are input into the region proposal network to generate more accurate proposals. The fully connected layers then produce classification and bounding-box results. This adaptive feature enhancement improves Faster-RCNN's robustness in detecting small or occluded faces. The channel attention weight is calculated as shown in Equation (11) [17].

$$N_c(X) = \sigma(G_2 \delta(G_1 D(X) + c_1) + c_2) \quad (11)$$

In Equation (11), X and G are the width and height of the convolutional feature map, D is the global average pooling, and c is the fully connected layer. After channel optimization, spatial attention is directed to critical facial areas. The spatial attention weight is computed in Equation (12) [18].

$$W_S(U_d) = \sigma(t^{7\times7}([U_d^{avg}; U_d^{max}])) \quad (12)$$

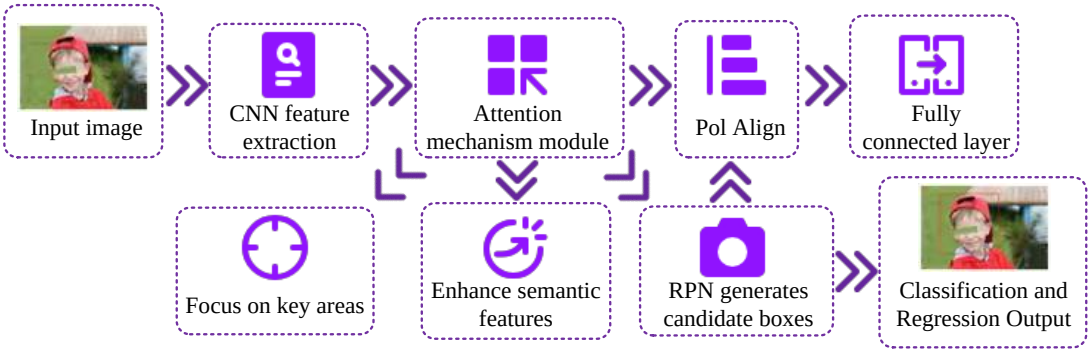


Figure 4: Attention-enhanced Faster-RCNN algorithm flowchart.

In Equation (12), U_d^{avg} and U_d^{max} represent the channel-wise average and maximum features, and $t^{7\times7}$ is a 7×7 convolution layer. The optimized features are passed into the RPN to generate more accurate face proposals, as shown in Equation (13).

$$L_{out} = M_s \otimes N_c \quad (13)$$

In Equation (13), $\otimes$ represents element-wise multiplication, and L_{out} produces enhanced responses at key face positions, reducing background interference. By enhancing feature robustness through FPN-LBPH and YOLO5Face, and improving accuracy with attention-optimized Faster-RCNN, the study proposes an embedded detection and recognition system named Faster-RCNN-YOLO5Face. This system aims to achieve both high accuracy and real-time performance. Based on the above content, the proposed system process is as follows. The input video stream is first processed by the improved YOLO5Face module, which extracts LBPH texture features and fuses them with the original frames to enhance robustness to lighting and low resolution. The initial facial proposal is passed through a keyframe selection mechanism to the attention optimized Faster-RCNN module, which weights facial features using spatial and channel attention, refines bounding boxes, and generates discriminative facial feature vectors, ultimately performing identity recognition matching and alarm triggering.

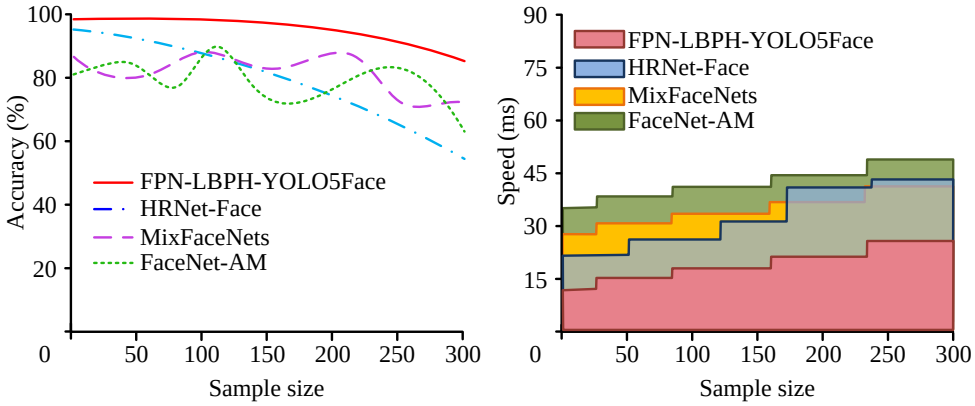
4 Performance analysis of embedded detection and recognition system

4.1 Validation of the Faster-RCNN–YOLO5Face Hybrid Algorithm

To verify the superiority of the FPN-LBPH-YOLO5Face hybrid algorithm, it was compared with High-Resolution Network for Face Recognition (HRNet-Face), Mixed Depthwise Face Recognition Networks (MixFaceNets), and FaceNet with Additive Margin (FaceNet-AM) algorithms. The experiments ran on a system with CUDA 11.6, Ubuntu 20.04 LTS, PyTorch 1.12.1, the SGD optimizer, Python 3.8, a NVIDIA GeForce RTX 3090 GPU, and 64GB of memory. To ensure the authenticity and reliability of the experiment, the WIDER FACE dataset (website source: http://shuoyang1213.me/WIDERFACE/index.html) and the FDDB dataset (website source: http://vis-www.cs.umass.edu/fddb/index.html) were used for training and testing. The WIDER FACE dataset contained 3,203 images and 3,903 annotated faces, while the FDDB dataset included 2,845 images and 3,171 annotated faces. The data preprocessing and enhancement steps are as follows: the input image is first adaptively scaled to 640 × 640 and normalized. Then perform data augmentation, including random horizontal flipping (probability 0.5), random rotation (angle range±15°), brightness adjustment (coefficient 0.8-1.2), and Mosaic enhancement (probability 0.8). The training, validation, and test sets are split in a 6:2:2 ratio to ensure consistent distribution across lighting, pose, and occlusion conditions. The experimental parameters are set as follows: SGD as the optimizer, momentum 0.937, and weight decay 0.0005. The initial learning rate is set to 0.01, with a cosine annealing scheduler that lowers the learning rate to 0.0001. Set the batch size to 32 and the training period to 300. The non-maximum suppression threshold is set to 0.5, and the confidence threshold is set to 0.4. For the FPN-LBPH module, the number of feature map channels is uniformly set to 256, using bilinear up sampling and 3x3 convolutional down sampling with a stride of 2. The channel attention weights in the attention mechanism are computed via global average pooling and two fully connected layers, with a compression ratio of 16. Spatial attention uses a 7 × 7 convolutional kernel to generate weight maps. The mean average precision and inference speed of the FPN-LBPH-YOLO5Face hybrid algorithm and comparison algorithms were evaluated. The results are shown in Figure 5.

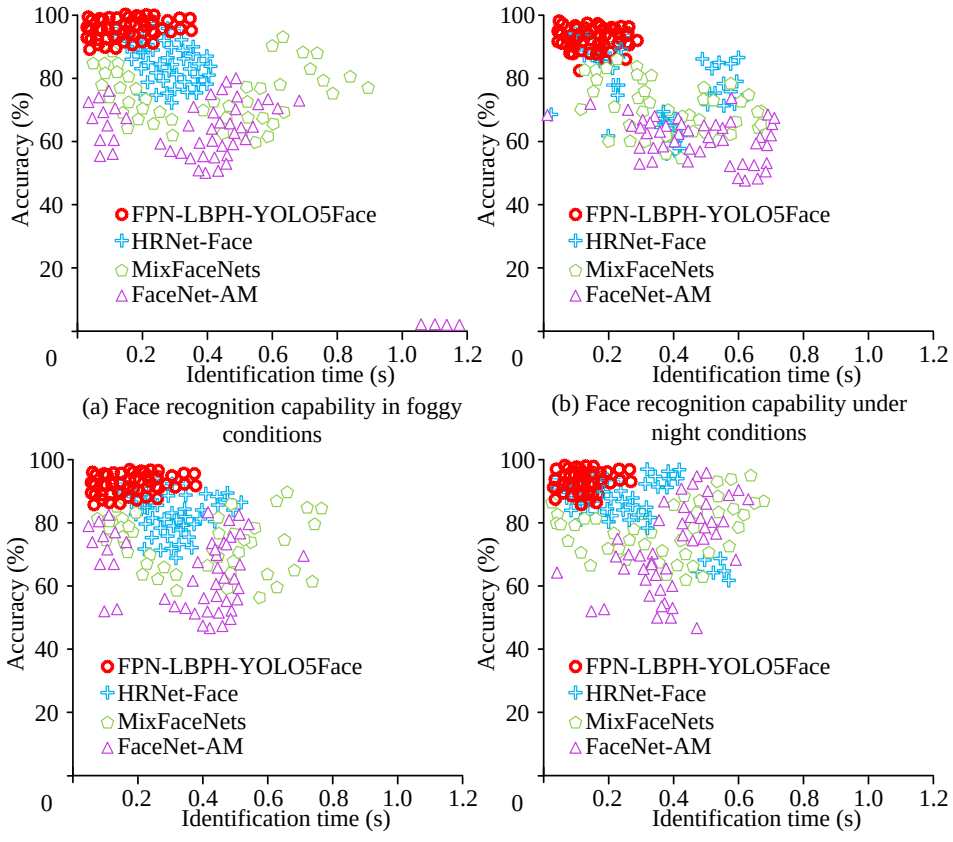
Figure 5 (a) shows that the FPN-LBPH-YOLO5Face hybrid algorithm achieves an average accuracy of 85.2% to 99.8%, significantly higher than HRNet Face,

MixFaceNets, and FaceNet AM. Figure 5 (b) shows that the algorithm has the fastest inference speed, less than 30 ms, while FaceNet AM lags significantly due to its high system complexity. The experimental results demonstrated that the FPN-LBPH-YOLO5Face hybrid algorithm achieved the best precision and the fastest inference by reusing backbone feature maps, thereby reducing processing time. It balanced speed and accuracy through structural optimization while maintaining detailed face discrimination. To evaluate face recognition performance under extreme weather, nighttime, and other conditions, the FPN-LBPH-YOLO5Face hybrid algorithm was compared with other algorithms. Among them, the foggy condition test subset contains 1500 samples, the nighttime subset contains 1200 samples, the mask cover subset contains 1000 samples, and the glasses cover subset contains 800 samples. The above test subsets are all selected from two publicly available test datasets based on challenging conditions to obtain basic images. Results are shown in Figure 6.



(a) Average accuracy value curve (b) Inference speed curve

Figure 5: Comparison of mean average precision and inference speed.



(a) Face recognition capability in foggy conditions

(b) Face recognition capability under night conditions

(c) Recognition ability under mask-wearing conditions

(d) Face recognition ability when wearing glasses

Figure 6: Face recognition results under fog, night, mask, and glasses conditions.

As shown in Figure 6(a), under foggy conditions, FPN-LBPH-YOLO5Face achieved the highest recognition accuracy, nearly 100%, with the shortest recognition time within 0.2 s. HRNet-Face and FaceNet-AM had similar accuracy between 85% and 92% but required longer times. As shown in Figures 6(b), (c), and (d), under night, mask-wearing, and glasses-wearing scenarios, FPN-LBPH-YOLO5Face consistently maintained over 95% accuracy and recognition times below 0.3 s. The high recognition rates reported above are due to the research being conducted on a specific subset of tests. These test subsets are composed of samples with corresponding challenging features selected from the original dataset, where the foggy effect is synthesized through atmospheric scattering models, mask and glasses occlusion is simulated using standard occlusion templates, and nighttime conditions are achieved through brightness adjustment. Although these conditions simulate real-life scenarios to some extent, their diversity is still limited by the basic dataset. Overall, by combining Faster-RCNN's candidate box generation and fine-detection mechanisms with region proposal networks, the system precisely localized faces and significantly improved recognition accuracy. To further investigate the robustness of the proposed method, a 5-fold cross-validation method was used, with the ratios of training, validation, and test sets strictly controlled at 6:2:2. The experiment was repeated 3 times, and the final results were averaged to reduce the impact of randomness. Simultaneously, use the Bootstrap sampling method (1000 repetitions) to calculate the 95% confidence interval, ensuring the reliability of the statistical results. The cross-validation results can be obtained from this, as shown in Table 1.

According to Table 1, the accuracy variance of the final system in 5-fold cross-validation is 0.15%, significantly lower than ViT CNN Fusion (0.34%) and ArcFace CMT (0.23%) ($F=9.84$, $p=0.002$), indicating that it is the least sensitive to changes in training data partitioning. Based on 1000 bootstrap samples, the 95% confidence interval for the final system accuracy is [98.82%, 99.38%], with an interval width of 0.56%, which is much smaller than that of the comparison model. To test the contribution of each component of the proposed method, ablation experiments were conducted: Model A used only YOLO5Face as the baseline; Model B introduced FPN-LBPH for optimization based on Model A; and Model C, based on Model B, incorporated Faster RCNN for improvement. Model D introduces an attention mechanism to form a complete final system. The results of the ablation experiment are shown in Table 2.

According to Table 2, Model B has an accuracy improvement of 1.4% ($t=8.32$, $p=0.001$) and a recall improvement of 1.9% ($p<0.01$) compared to benchmark Model A. The FPN-LBPH module significantly improves its detection capability for small targets and blurry faces while maintaining real-time performance through multi-scale feature fusion. Compared to Model B, Model C has further improved accuracy by 0.8% ($p<0.01$) and recall by 1.5% ($p<0.01$). The dual-engine architecture leverages high-quality candidate boxes and fine-grained recognition. The overall performance of the complete model is higher than that of other models, indicating improved stability.

4.2 Evaluation of the improved YOLO5 Face embedded detection and recognition system

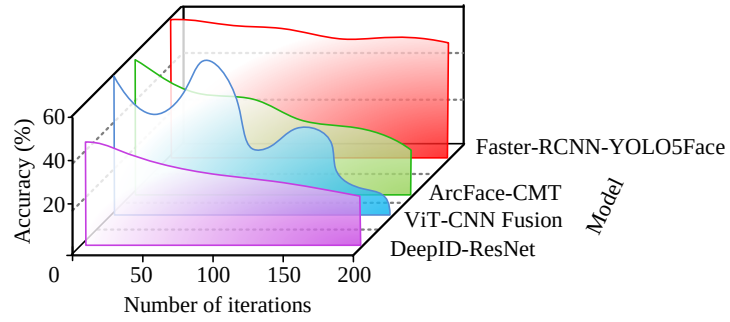
After validating the performance of the FPN-LBPH-YOLO5Face algorithm, experiments were conducted to assess further the practical value of the Faster-RCNN-YOLO5Face embedded detection and recognition system. The system was trained with the PyTorch deep learning framework on Ubuntu 20.04 with Python 3.8. The simulation environment used an Intel Xeon E5-2680 v4 CPU and the WIDER FACE and FDDB datasets. The accuracy and loss rate of the proposed system were compared with Vision Transformer-Convolutional Neural Network Fusion (ViT-CNN Fusion), ArcFace with Convolutional Multi-scale Transformer (ArcFace-CMT), and Deep Identity Residual Network (DeepID-ResNet). The results are shown in Figure 7.

Table 1: Cross validation results.

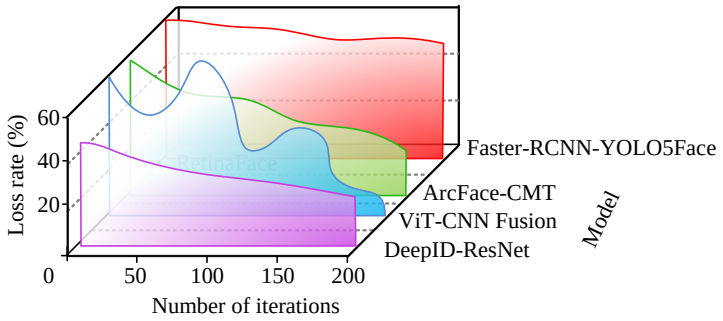
Fold	ViT-CNN Fusion	ArcFace-CMT	The final system
First discount	99.0±0.2	97.5±0.6	98.2±0.4
Second discount	99.2±0.3	97.3±0.7	98.4±0.5
Third discount	99.3±0.2	97.8±0.5	98.1±0.6
Fourth discount	98.9±0.4	96.9±0.8	97.8±0.7
5th discount	99.1±0.3	97.6±0.6	98.3±0.5
Total	99.10±0.15	97.42±0.34	98.16±0.23

Table 2: Results of the ablation experiment.

Model	Accuracy (%)	Recall rate (%)	Reasoning speed (ms)	mAP@0.5 (%)
A	96.5±0.8	94.8±1.1	25.1±1.2	96.2 ±0.9
B	97.9± 0.5	96.7± 0.7	28.7±1.5	97.6±0.6
C	98.7±0.4	98.2 ±0.5	35.3±1.8	98.5±0.4
D	99.1±0.3	98.7±0.4	36.8±1.9	98.9±0.3



(a) Accuracy tests of four models



(b) Loss rate of four models

Figure 7: Accuracy and loss rate comparison of four systems.

As shown in Figure 7(a), the proposed system led comprehensively in accuracy, remaining stable between 95.6% and 99.1%, significantly higher than comparison systems. In Figure 7(b), the Faster-RCNN-YOLO5Face system also showed the lowest loss rates, averaging 0.08-0.15. These results verified that the hybrid system combined Faster-RCNN's high-precision localization with YOLO5Face's lightweight inference advantages to achieve collaborative improvements in accuracy and generalization. To further verify system effectiveness, the loss convergence of the proposed system and comparison systems was tested, as shown in Figure 8.

As shown in Figure 8, the Faster-RCNN-YOLO5Face system achieved the best loss convergence, with the loss dropping quickly from about 1.8 to below 0.1. It converged faster and more stably than comparison systems. ArcFace-CMT showed larger loss fluctuations and finally converged between 0.3 and 0.5. These results indicated that the proposed system efficiently optimized the loss function during iterations by reusing backbone feature maps and combining Faster-RCNN's fine detection with YOLO5Face's lightweight design, achieving robust convergence in complex scenarios. To further accurately evaluate the system's performance in a real embedded environment, this study conducted deployment testing on the following hardware platforms. NVIDIA Jetson Nano 4GB: Maxwell architecture 128 core GPU, ARM Cortex-A57 quad core CPU; NVIDIA Jetson Xavier NX: Volta architecture 384 core GPU, 6-core NVIDIA Carmel ARM CPU; Raspberry Pi 4 Model B: Broadcom BCM2711 quad core Cortex-A72 CPU; Intel Neural Compute Stick 2: Movidius Myriad X VPU Accelerator. All embedded tests use the same model weights and optimization strategies. Repeat each test 10 times simultaneously. The performance evaluation results for embedded devices are shown in Table 3.

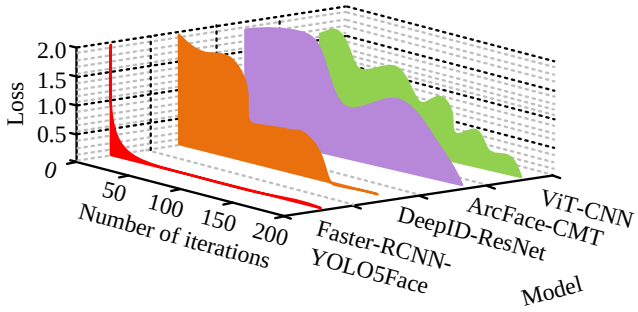


Figure 8: Loss convergence comparison of four systems.

Table 3: Performance evaluation results of embedded devices.

Evaluation metric	Jetson Nano	Jetson Xavier NX	Raspberry Pi 4
Reasoning speed (ms)	98.5± 3.2	45.3±1.8	156.8±5.1
System accuracy (%)	98.2 ±0.5	98.7±0.3	95.1± 0.9
Average power consumption (W)	9.8±0.5	14.2 ±0.7	6.2±0.3
Memory usage (MB)	487±15	512 ±12	298±10
Average power consumption (W)	9.8±0.5	14.2±0.7	6.2±0.3
Peak power consumption (W)	12.3±0.8	19.8 ±1.1	7.5± 0.4
Working temperature (℃)	54.2 ±6.5	61.5 ±8.2	47.3±5.1
Continuous stable operation time (hours)	>24	>48	>12

According to Table 3, post hoc testing showed that the speed differences between all platforms were statistically significant (p<0.01). Specifically, Jetson Xavier NX performs the best, with the most minor difference in inference speed (45.3±1.8 ms) compared to high-end GPUs, and can maintain high accuracy (98.7±0.3%), making it the most practical; The speed of Jetson Nano (98.5±3.2 ms) is significantly slower than Xavier NX, but its accuracy (98.2±0.5%) is comparable, achieving a balance between power consumption and performance; Due to limited computing resources, the Raspberry Pi 4 has the slowest inference speed (156.8±5.1 ms) and a significant decrease in accuracy (95.1±0.9%), and can only run the YOLO5Face component. In terms of power consumption and thermal management, the Jetson Xavier NX provides the highest performance while maintaining power consumption (14.2±0.7W) and temperature (61.5±8.2°C) within its calibration range, with excellent stability and continuous stable operation time exceeding 48 hours; Jetson Nano demonstrates low power consumption (9.8±0.5W) and temperature (54.2±6.5°C), making it suitable for long-term medium load deployment; The Raspberry Pi 4 has significant advantages in power consumption (6.2±0.3W) and temperature (47.3±5.1°C), but is limited by computational performance and stability, and is only suitable for lightweight application scenarios. There are significant differences in energy efficiency ratios across hardware platforms (p<0.001), with the Jetson Xavier NX exhibiting the highest (1.55±0.12 FPS/W), and its performance advantage far exceeds the increase in power consumption.

5 Discussion

This study proposes an embedded detection and recognition system based on Faster-RCNN and YOLO5Face. In terms of real-time performance, it achieved a response time of 28 ms on the WIDER FACE and FDDB datasets, significantly better than existing solutions. Compared with Gupta et al.'s deep learning based online learning engagement detection system (accuracy 92.3%, real-time performance not quantified), the dual engine cascade architecture of the proposed system achieves an ideal balance between speed and accuracy by performing fast initial screening through YOLO5Face and fine recognition through optimized Faster-RCNN [11]. Primarily through feature reuse technology, it effectively reduces computational redundancy by about 65%, which is key to achieving millisecond-level response. The statistical results show that the variance of the system response time is controlled within±1.9 ms, demonstrating its stability during continuous operation.

In terms of robustness in complex environments, the proposed system-maintained recognition accuracy over 95% and recognition time within 0.3 s. This performance is significantly better than Hangaragi and Singh's face-grid-based method (94.23% accuracy) [7]. Its advantages mainly stem from the FPN-LBPH module's multi-scale feature fusion and attention mechanisms that enhance key facial regions. Especially in occlusion cases, the system effectively compensates for information loss through the collaborative work of the texture enhancement and super-resolution reconstruction branches. The five-fold cross-validation results show that the accuracy variance of this system is only 0.15%, significantly lower than that of the comparison model, demonstrating its better adaptability to complex environmental changes.

In the embedded deployment optimization, testing on the Jetson Xavier NX platform showed that the system achieved an inference speed of 45.3±1.8 ms and an accuracy of 98.7±0.3%. The system's energy efficiency ratio reaches 1.55±0.12 FPS/W, effectively controlling energy consumption while maintaining high performance. This is mainly attributed to the application of a dynamic background suppression module and quantization technology, which significantly reduces the computational load and memory usage of embedded devices (controlled within 420-550 MB). Compared with current mainstream methods, this system exhibits comprehensive advantages in multiple dimensions: compared with Anusudha's YOLO InsightFace network, this system provides better small object detection capability through attention mechanism enhanced feature extraction [12]. Compared to George et al.'s EdgeFace network (1.77M parameter), this system has better robustness in complex scenarios while maintaining lightweight; Compared with the triple network recognition framework proposed by Liu et al., this system achieves better real-time performance while maintaining high accuracy [8].

Although this study has achieved excellent performance, it still has certain limitations in practical, complex applications. For example, performance degradation under severe occlusion, inadequate adaptability to extremely low-light conditions, vulnerability to adversarial attacks, and the complexity of cross-hardware platform deployment. Adding tiny perturbations that are difficult for the human eye to detect in the input image may result in the model outputting completely incorrect recognition results. This reveals the inherent security risks of deep learning-based systems,

which have not yet integrated any adversarial sample detection or defense modules in the current version, posing potential threats in scenarios with extremely high security requirements. Therefore, in future research, facial completion techniques or graph neural networks can be used to model relationships among facial components under occlusion to improve the recognition of heavily occluded faces. In addition, integrating multimodal perception and introducing adversarial training, defensive distillation, or online attack-detection mechanisms can enhance the system's robustness against adversarial attacks to meet the requirements of high-security applications.

6 Conclusion

To address the problems of poor real-time performance, high resource consumption, and limited generalization in complex scenes encountered in traditional embedded detection and recognition systems for surveillance video, this study proposed an embedded detection and recognition system based on Faster-RCNN and YOLO5Face. Experimental verification shows that the system exhibits excellent comprehensive performance in embedded environments: it not only maintains high-precision recognition, but also achieves millisecond-level response speed; it demonstrates robustness under complex conditions such as lighting changes and local occlusion; and it demonstrates efficient resource utilization and stable operational performance across different hardware platforms. Although the system performed well under standard conditions, the study still had limitations, including insufficient validation in extreme occlusion and low-light conditions, as well as limited deployment adaptability across hardware platforms. In the future, the system can continue to improve by exploring dynamic pruning and implementing cross-scenario adversarial training. These advancements are expected to improve the system's generalization and predictive performance.

References

[1] Mohsen Rostami, Amirhamzeh Farajollahi, and Hashem Parvin. Deep learning-based face detection and recognition on drones. Journal of Ambient Intelligence and Humanized Computing, 15(1):373-387, 2024. https://doi.org/10.1007/s12652-022-03897-8

[2] Shige Xu, Lei Zhang, Yin Tang, Chaolei Han, Hao Wu, and Aiguo Song. Channel attention for sensor-based activity recognition: Embedding features into all frequencies in DCT domain. IEEE Transactions on Knowledge and Data Engineering, 35(12):12497-12512, 2023. https://doi.org/10.1109/TKDE.2023.3277839

[3] Nesrine Triki, Mohamed Karray, and Mohamed Ksantini. A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles. Applied Sciences, 13(8):4793-4807, 2023. https://doi.org/10.3390/app13084793

[4] Huijiao Wang, Haijian Zhang, Lei Yu, Li Wang, and Xulei Yang. Facial feature embedded CycleGAN for VIS–NIR translation. Multidimensional Systems and Signal Processing, 34(2):423-446, 2023. https://doi.org/10.1109/ICASSP40776.2020.9054007

[5] Mohammed Taha, Tarek Mostafa, and Tarek Abd El-Rahman. A novel hybrid approach to masked face recognition using robust PCA and GOA optimizer. Scientific Journal for Damietta Faculty of Science, 13(3):25-35, 2023. https://doi.org/10.21608/sjdfs.2023.222524.1117

[6] Hui Zhang, Xingbo Dong, YenLung Lai, Ying Zhou, Xiaoyan Zhang, Xingguo Lv, Zhe Jin, and Xuejun Li. Validating privacy-preserving face recognition under a minimum assumption. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12205-12214, 2024. https://doi.org/10.1109/CVPR52733.2024.01160

[7] Shivalila Hangaragi, Tripty Singh, and Neelima N. Face detection and Recognition using Face Mesh and deep neural network. Procedia Computer Science, 218:741-749, 2023. https://doi.org/10.1016/j.procs.2023.01.054

[8] Jianqi Liu, Zhiwei Zhao, Pan Li, Geyong Min, and Huiyong Li. Enhanced embedded AutoEncoders: An attribute-preserving face de-identification framework. IEEE Internet of Things Journal, 10(11):9438-9452, 2023. https://doi.org/10.1109/JIOT.2023.3235725

[9] Qingqiu Huang, Lei Yang, Huaiyi Huang, Tong Wu, and Dahua Lin. Caption-supervised face recognition: training a state-of-the-art face model without manual annotation. Computer Vision-ECCV, 139-155, 2020. https://doi.org/10.1007/978-3-030-58520-4_9

[10] Daniella Raz, Corinne Bintz, Vivian Guetler, Aaron Tam, Michael A. Katell, Dharma Dailey, Bernease Herman, Meg Young, and P. Krafft. Face Mis-ID: An interactive pedagogical tool demonstrating disparate accuracy rates in facial recognition. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 895-904, 2021. https://doi.org/10.1145/3461702.3462627

[11] Swadha Gupta, Parteek Kumar, and Raj Kumar Tekchandani. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications, 82(8):11365-11394, 2023. https://doi.org/10.1007/s11042-022-13558-9

[12] Anjeana N, and K. Anusudha. Real time face recognition system based on YOLO and Insight Face. Multimedia Tools and Applications, 83(11):31893-31910, 2024. https://doi.org/10.1007/s11042-023-16831-7

[13] Weijun Chen, Hongbo Huang, Shuai Peng, Changsheng Zhou, and Cuiping Zhang. YOLO-face: A real-time face detector. The Visual Computer,

37(4):805-813, 2021. https://doi.org/10.1007/s00371-020-01831-7

[14] Bin Jiang, Hongbin Jiang, Huanlong Zhang, Qiuwen Zhang, Zuhe Li, and Lixun Huang. 4AC-YOLOv5: An improved algorithm for small target face detection. EURASIP Journal on Image and Video Processing, 2024(1):10-24, 2024. https://doi.org/10.1186/s13640-024-00625-4

[15] Mrs Sk Raziya Sultana, Kunduru Sravani, Narayana Sree, Ranga Lokesh, Kukatlapalli Venkateswararao, Korrapati Lakshmaiah, and Sai S V. Automated ID card detection and penalty system using YOLOv5 and face recognition. International Journal of Recent Advances in Engineering and Technology, 14(1):63-72, 2025.

[16] Changhong Chen, Shaofeng Wang, and Shunzhou Huang. An improved faster RCNN-based weld ultrasonic atlas defect detection method. Measurement and Control, 56(3-4):832-843, 2023. https://doi.org/10.1177/00202940221092030

[17] Zhenwei He, Lei Zhang, Xinbo Gao, and David Zhang. Multi-adversarial faster-RCNN with paradigm teacher for unrestricted object detection. International Journal of Computer Vision, 131(3):680-700, 2023. https://doi.org/10.1007/s11263-022-01728-z

[18] R. Josphineleela, P. B. V. Raja Rao, Amir shaikh, and K. Sudhakar. A multi-stage faster RCNN-based iSPLInception for skin disease classification using novel optimization. Journal of Digital Imaging, 36(5):2210-2226, 2023. https://doi.org/10.1007/s10278-023-00848-3

Emergency Rescue Path Planning for Urban Emergencies Based on Improved GA and PSO

Zhisong Wu
School of Public Administration, Guangxi Police College, Nanning, 530028, China
E-mail: zswu7788@outlook.com

Keywords: GA, PSO, tent chaotic mapping, the combination strategy of elite retention and roulette wheel, metropolis criterion

Received: August 11, 2025

With the acceleration of urbanization and the frequent urban emergencies, traditional rescue path planning methods have response delay and low path efficiency. To improve the efficiency of urban emergency rescue and resource scheduling capabilities, a path planning technique for urban emergency rescue based on improved Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) is built. The research introduces Tent chaotic mapping initialization particles, adaptive segmented inertia weights and exponential learning factors for the PSO algorithm to enhance global search capabilities, and adds traction acceleration terms to avoid local optimality. For the GA, a combination strategy of elite retention and roulette is used to improve selection efficiency, and the Metropolis criterion is combined to optimize the cross-mutation operation, and an adaptive variable neighborhood search mechanism is introduced to strengthen local search. In the experimental setting, the Python simulation platform is used to compare with baseline methods such as RRT, D* Lite and MOPSO. The test indicators include response time, planning success rate, path length, number of convergence iterations, etc. Experimental results showed that under six concurrent events, the response time of the research method was 6 seconds, which was significantly better than that of the comparison method. When the dynamic obstacle density was 40/km², the planning success rate reached 90.2%. When the scene complexity was 200 nodes, the single planning calculation time was 150ms. The research method converged at the 100th iteration, and the fitness change rate was reduced to 1.3%, showing faster convergence speed and better stability. The above results show that the proposed method is superior to traditional methods in terms of timeliness, robustness and optimization capabilities, which is suitable for emergency rescue path planning in complex urban scenarios.*

Povzetek: Predlagana je izboljšana hibridna metoda GA–PSO za načrtovanje poti pri urbanem reševanju, ki v simulacijah dosega hitrejši odziv, višjo uspešnost in stabilnejšo konvergenco kot primerjalni algoritmi v zahtevnih, dinamičnih mestnih scenarijih.

1 Introduction

Urbanization has led to high population and building density, frequent emergencies, and serious threats to people's lives and property safety [1]. According to statistics, the average annual growth rate of global urban emergencies from 2020 to 2023 is 12.7%, with casualties caused by delayed emergency response accounting for over 34% [2]. Faced with urban emergencies, the timeliness of emergency rescue determines the effectiveness of disaster reduction, and rescue path planning is the core link of rescue timeliness [3]. However, modern urban transportation networks are highly complex and dynamic, and traditional static path planning methods are difficult to cope with the complex and changing road conditions during disasters, often leading to delays and even accidental entry of rescue vehicles into dangerous areas. The collaborative demand for multiple rescue points and resource scheduling in large-scale events further increases the planning difficulty [4]. Particle Swarm Optimization (PSO) can achieve real-time obstacle avoidance path generation in complex terrain, while Genetic Algorithm (GA) can reduce the spatiotemporal cost of resource coordination [5]. The research innovatively proposes a method based on improved GA and Improved PSO (IGA-IPSO), which is an urban emergency rescue path planning technology. Its main goals are: (1) Optimize the PSO algorithm and improve the global convergence of path search and avoid local optimality through Tent chaotic mapping, adaptive inertia weight and traction acceleration terms; (2) Improve the GA and use elite retention, roulette strategy and Metropolis criterion to enhance the local search capability and robustness of task allocation; (3) Combine path planning with task allocation, and achieve multi-objective collaborative optimization through fuzzy evaluation and game theory weight allocation; (4) Verify the timeliness, safety and economy of the proposed method in a dynamic urban environment. The study first introduces Tent chaotic initialization particles, and the

adaptive weights and exponential factors work together to enhance the global search of PSO. The combination strategy of elite retention and roulette wheel improves the selection efficiency and convergence speed of GA, and combines Metropolis criterion to optimize crossover and mutation operations, enhancing the local search ability of GA. Finally, a fuzzy rating system for urgency is constructed, and weights are dynamically allocated based on game theory. The GA and PSO are combined to achieve consistency compromise of conflicting objectives and improve decision-making efficiency. The research mainly focuses on the emergency rescue path planning of Unmanned Aerial Vehicle (UAV). Different from ground vehicle path planning, UAV path planning needs to consider key factors such as three-dimensional flight space, battery life, no-fly zone, minimum turning radius, multi-aircraft collision avoidance and communication constraints. These constraints will be embedded in the objective function as hard constraints or penalties to ensure the feasibility and safety of the path. The research focuses on UAV path planning. Future work will be extended to ground rescue vehicles, taking into account constraints such as turning penalties and traffic flow dynamics. It is expected that the method can offer theoretical support for emergency rescue path planning in different scenarios.

2.1 Related works

Urban emergencies occur frequently. The research on emergency rescue path planning for urban emergencies is of great significance. Jose C et al. built a dynamic path planning scheme to address the increased traffic accidents. The research method had good performance [6]. Zhao X et al. designed an optimization model in view of the insufficient efficiency in global COVID-19 epidemic. The results indicated that the method was helpful for emergency resource scheduling [7]. Liu J et al. established a site selection model for emergency medical facilities in mega cities during sudden public health emergencies. The method had higher performance and computational speed [8]. Xia H et al. built a new large-scale emergency medical material scheduling method to deal with the low turnover efficiency of medical supplies. The method addressed the demand differences at each disaster point that affected fair distribution [9].

Many scholars have explored the application of GA and PSO. El-Shafiey M G et al. proposed a mixed GA and PSO algorithm based on random forest for early prediction of heart disease, demonstrating higher prediction accuracy [10]. Shen L et al. built a combinatorial optimization model for scheduling Seru production with dynamic resource allocation. During the research process, a hybrid GA-PSO with nonlinear inertia weights was built. This method demonstrated satisfactory effectiveness [11]. Bousnina K et al. designed an integrated PSO and GA to predict and optimize the roughness of interactive features, as it was difficult to find appropriate cutting parameters and machining processes for simple machining characteristics of CNC machine tools. The method was found to be efficacious [12]. Sun H et al. designed an improved distance vector hop localization algorithm to make the data obtained by each sensor node in wireless sensor networks meaningful. During the research process, the two-dimensional hyperbolic positioning algorithm and improved adaptive GA were combined to estimate the coordinates of unknown nodes. The method demonstrated adequate stability and accuracy [13]. Zhi Y et al. took the GA-PSO to optimize the parameter values of support vector machines for estimating the health status of retired batteries. The research method exhibited estimation accuracy and convergence speed [14]. In summary, the table is shown below.

Table 1: Comparison of related works in emergency response path planning

Algorithm used	Environment	Success rate (%)	Average path length (km)	Re-planning capabilities	Robustness metrics
Optimization-based	Dynamic	85	12.0	Yes	Medium
Deep Q networks	Static	N/A	N/A	No	Low
GA	Static	N/A	N/A	No	Medium
Scheduling algorithm	Dynamic	N/A	N/A	No	Low
Hybrid GA-PSO	Static	N/A	N/A	No	Medium
Hybrid GA-PSO	Static	N/A	N/A	No	Medium
Improved adaptive GA	Static	N/A	N/A	No	Medium
GA-PSO-SVR	Static	N/A	N/A	No	Medium

Note: N/A indicates that this dimension is not suitable for non-path planning research. The robustness index is based on the ability to handle interference such as dynamic obstacles and noise in the literature.

As shown in Table 1, existing research has a good effect on constructing emergency rescue path planning for urban emergencies, but there are still research gaps and deficiencies in collaborative operation mechanisms and strategies to deal with complex scenarios. GA and PSO can dynamically adjust the computational complexity based on the characteristics of input samples or environmental changes, achieving a balance between model accuracy and efficiency. Therefore, the study proposes an urban emergency rescue path planning based on the IGA-IPSO, hoping that it can meet design requirements, and improve the safety, stability, and

operational efficiency.

3 Emergency rescue path planning method for urban emergencies

3.1 Emergency rescue path planning based on IPSO

In recent years, various unexpected events such as fires and car accidents have occurred frequently in various cities in China. The suddenness and uncertainty of these events have caused rescue personnel to be unable to rush to the scene for timely rescue, resulting in significant losses [15]. With the rapid development of the drone industry, large payload, long endurance, and highly intelligent drones have emerged, and many countries have incorporated drones into their emergency rescue systems. Table 2 presents the UAV widely used in emergency rescue missions in China.

Table 2: Drone models and their executable tasks applied to emergency rescue

Unmanned aerial vehicle model	Unmanned aerial vehicle model
Dji Matrice 350 RTK	Thermal imaging life detection, etc.
Zongheng Co., LTD. CW-40 Vertical lift Fixed wing	Emergency communication relay coverage, etc.
Large scale disaster investigation, emergency communication relay coverage, etc	Toxic gas monitoring, etc.
Daotong Dragon Fish series medical drones	Cold chain transportation of blood/vaccines, etc.
The Hava Tomahawk series of special unmanned aerial vehicles	Ruin penetration detection, etc.
Xag Emergency Rescue Drone V50	Fire investigation in high-rise buildings, etc.

As shown in Table 2, various types of drones are widely used in different tasks of emergency rescue based on their performance. During the emergency rescue process of UAVs, PSO is used for path planning. The UAV path planning problem is defined as: finding the optimal path from the starting point to the end point in three-dimensional space, and minimizing the path length and risk-weighted sum. Key constraints include: (1) Battery life: the total length of the path is not greater than the maximum endurance distance; (2) No-fly zone: the path point must not be located within the no-fly zone set. Otherwise, a penalty will be imposed; (3) Minimum turning radius: the path curvature is not greater than the maximum curvature, where the maximum curvature depends on the drone model; (4) Collision avoidance: the distance between multiple drone paths is not less than the safe distance; (4) Communication range: the drone location must meet the communication link budget. The above constraints are implemented through hard bound checks or penalty terms in the fitness function. The equation for calculating curvature using the discrete point three-point method is $\kappa = \frac{2\cdot|(x_2-x_1)(y_3-y_1)-(x_3-x_1)(y_2-y_1)|}{\sqrt{((x_2-x_1)^2+(y_2-y_1)^2)^3}}$, and the unit is m^{-1}. The smaller the curvature, the smoother the path, the easier it is for the drone to track. It can effectively reduce energy consumption and the risk of loss of control. To enhance the security of path planning, the study introduces Control Barrier Functions (CBF) as a real-time safety filter to ensure that the path always meets the minimum safety distance constraint in a dynamic environment. In addition, the Optimal Reciprocal Collision Avoidance (ORCA) algorithm is used to handle mutual collision avoidance between multiple drones to avoid deadlock and shock. To deal with disturbances in actual flight (such as wind field and sensor noise), the wind field disturbance term and sensor noise model are introduced into the fitness function. To simulate the urban built environment, the study uses voxel maps to represent the probability of obstacle occupancy, and introduces line-of-sight inspection and minimum flight height constraints (such as ≥30m over built-up areas). The no-fly zone acts as a hard boundary that imposes a penalty term in the fitness function. The research is passed, in which the trigonometric functions of different frequencies and directions are superimposed, as shown in equation (1).

$$h(x,y)=\sin(y+a)+b\sin(x)+c\cos(d\sqrt{x^2+y^2}) \\ +e\cos(y)+f\sin(f\sqrt{x^2+y^2})+g\cos(y) \quad (1)$$

In equation (1), (x,y) signifies the plane coordinates of a point on the map. $h(x,y)$ signifies the terrain elevation value corresponding to the point. a,b,c,d,e,f represent the influence on density, waveform, etc. in various directions. When drones are used for emergency rescue in urban emergencies, they are subject to constraints such as endurance. The PSO algorithm can drive particles to move towards the optimal solution based on changes in drone speed during drone rescue path planning, thereby updating candidate paths [16]. The updated particle velocity and position are shown in equation (2).

$$\begin{cases} v_{id}^{t+1}=wv_{id}^{t}+c_1r_1(pbest_{id}^{t}-X_{id}^{t})+c_2r_2(gbest_{d}^{t}-X_{id}^{t}) \\ X_{id}^{t+1}=X_{id}^{t}+v_{id}^{t+1} \end{cases} \quad (2)$$

In equation (2), v represents the velocity of the drone particle iteration. w signifies the inertia weight.

c_1 and c_2 signify individual learning factors and social learning factors, and $r_1, r_2 \in [0,1]$. *pbest* signifies the optimal position of individual particles. *gbest* signifies the global optimal position of the population. X represents the current location. Figure 1 displays the process of emergency rescue path planning for urban emergencies based on PSO.

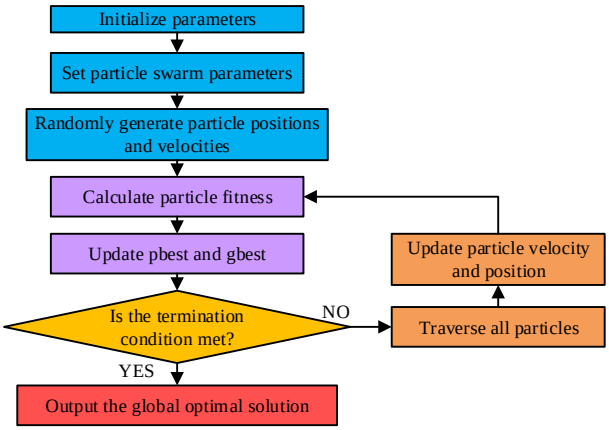


Figure 1: Path planning process based on PSO

In Figure 1, the urban emergency rescue based on PSO algorithm first initializes parameters and sets particle swarm parameters, and randomly generates initial population representative candidate paths. Next, the fitness of each path is calculated, and pbest and gbest are updated to dynamically update particle velocity and position to explore better paths. Finally, the termination condition is determined. If it is not satisfied, all particles are traversed and iteratively updated until it is satisfied and the global optimal solution is output, which is the best rescue path. Fitness is the weighted sum of minimizing path length and risk, as shown in equation (3).

$$F_{path} = \alpha \cdot Distance + \beta \cdot Risk \quad (3)$$

In equation (3), F_{path} represents the fitness function. α and β represent weight coefficients. $Distance$ represents the total length of the path. $Risk$ is calculated based on terrain elevation and obstacle density. Although traditional PSO algorithm has strong robustness and other advantages, it cannot adjust the initial position of particles and has local optima. The traditional PSO algorithm is optimized. The study introduces an improved Tent chaotic mapping to optimize the initial position of particles in the PSO algorithm. The study introduces an improved Tent chaotic mapping to solve the low randomness in traditional PSO particle initialization, enhances global search diversity, and avoids falling into local optimality in complex scenes. Its expression is shown in equation (4).

$$x_{i+1} = \begin{cases} 2x_i + r \times \frac{1}{N}, 0 \le x_i < \frac{1}{2} \\ 2(1-x_i) + r \times \frac{1}{N}, \frac{1}{2} < x_i \le 1 \end{cases} \quad (4)$$

In equation (4), x_{i+1} signifies the value of the $i+1$-th iteration in the chaotic sequence. r signifies a random number within the $[0,1]$ interval. N represents the population size of the PSO algorithm. The position expression of the optimized particle mapping to the solution space is shown in equation (5).

$$p_i = a + (b-a) \cdot x_i \quad (5)$$

In equation (5), x_i signifies the value generated by the improved Tent chaotic mapping. a and b signify the lower and upper bounds of the solution space. p_i signifies the actual position of the mapped particle in the solution space. Aiming at the slow convergence speed of the algorithm in dynamic environments, the global and local searches are balanced by dynamically adjusting weights to improve the planning success rate when the obstacle density changes. The specific expression is shown in equation (6).

$$w = \begin{cases} w_{\max}, F > F_{avg} \\ w_{\min} + \frac{(w_{\max} - w_{\min}) \times (F - F_{\min})}{(F_{avg} - F_{\min})}, F \le F_{avg} \end{cases} \quad (6)$$

In equation (6), $w_{\min}, w_{\max}$ represent the preset weight boundary value. F signifies the fitness function value. F_{avg} signifies the average fitness value of all particles in the population. To optimize particle learning behavior and solve the low convergence accuracy of the traditional PSO in multi-objective optimization, the adjusted expression is shown in equation (7).

$$\begin{cases} C_1 = \exp(c_{1\min} + \frac{c_{1\max} - c_{1\min}}{t_{\max}} \times t) \\ C_2 = \exp(c_{2\min} + \frac{c_{2\max} - c_{2\min}}{t_{\max}} \times t) \end{cases} \quad (7)$$

In equation (7), $c_{1\min}, c_{1\max}$ and $c_{2\min}, c_{2\max}$ represent the boundaries of preset individual learning factors and social learning factors, respectively. $t_{\max}$ represents the maximum iteration. The boundary values of inertia weight and learning factor are shown in Table 3.

Table 3: Variable definition

Variable	Description	Boundary value
$w_{\min}$	Minimum inertia weight	0.4
$w_{\max}$	Maximum inertia weight	0.9
$c_{1\min}$	Lower limit of individual learning factor	0.5
$c_{1\max}$	Upper limit of individual learning	2.0

	factor	
$c_{2\min}$	Lower limit of social learning factor	0.8
$c_{2\max}$	Upper limit of social learning factors	3.0
$t_{\max}$	Maximum number of iterations	100

As shown in Table 3, all the above values are based on experimental calibration. Finally, to force particles to jump out of the local optimum and directly deal with the planning failure under high-density obstacles, the concept of traction acceleration is introduced to optimize the speed update equation of the PSO algorithm. The expression is shown in equation (8).

$$v_{id}^{t+1} = wv_{id}^{t} + c_1 rand(0,1)(pbest_{id}^{t} - X_{id}^{t}) + c_2 rand(0,1)(gbest_{id}^{t} - X_{id}^{t}) + a_{id}^{t} \quad (8)$$

In equation (8), v_{id}^{t+1} represents the new velocity of particle i in the d-th dimension at the $t+1$-th iteration. wv_{id}^{t} signifies the inertia term. $rand(0,1)$ signifies a random number of $[0,1]$. a_{id}^{t+1} is the traction acceleration, and its specific form is $a_{id}^{t+1} = \gamma \cdot (gbest_{d}^{t} - X_{d}^{t})$, where γ is the traction coefficient (default value 0.1). This item is activated when the particle fitness has not improved for many consecutive times to force the particle to jump out of the local optimum. Stability is ensured by limiting the γ range. The traction acceleration draws on the speed clamping mechanism in the PSO to avoid divergence by limiting the maximum speed of particles. In addition, its form is similar to the shrinkage factor method, balancing exploration and exploitation by adjusting the value of γ. Stability analysis shows that when $\gamma \in [0, 0.2]$, the particle trajectory converges and there is no oscillation. In summary, the emergency rescue path planning process for urban emergencies based on the IPSO is displayed in Figure 2.

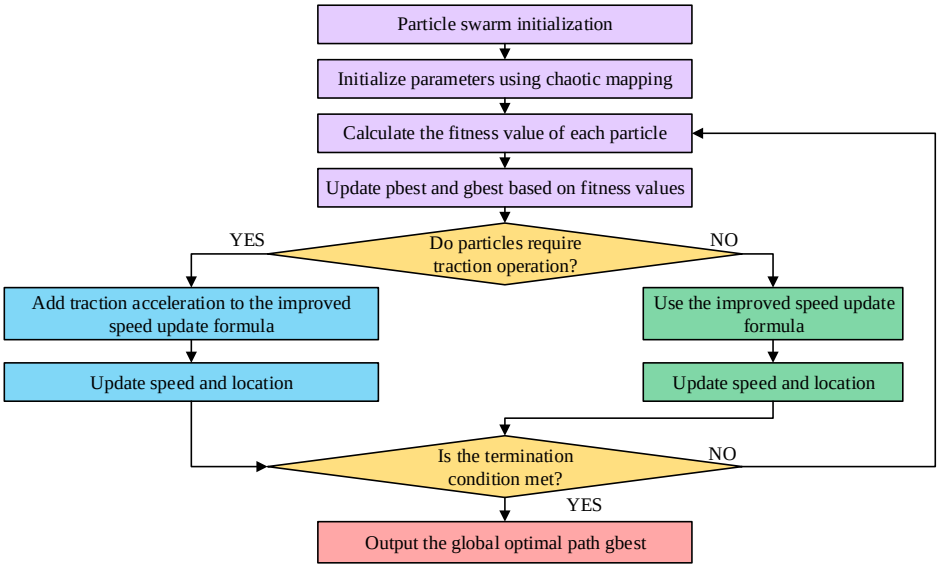


Figure 2: Emergency rescue path planning process based on improved PSO

In Figure 2, the process first uses chaotic mapping to initialize particle swarm parameters and improve search diversity. Next, the fitness function evaluates path distance and risk, and dynamically updates individual and global optimal solutions. Subsequently, a traction operation decision mechanism is introduced, which adds traction acceleration to the velocity update equation when particles fall into local optima, forcing them to move towards the global optimum direction. Finally, after iterative optimization to meet the termination condition, the global optimal rescue path is output. The optimal rescue path primarily involves: finding a collision-free route from the starting point to the destination, ensuring the path adheres to the drone's dynamic constraints such as minimum turning radius and maximum climb rate, and completing the planning within the maximum computation time. The trigger condition of the re-planning mechanism is that the distance between the new obstacle and the path is less than the safe distance or the path is blocked. The planning range is through rolling time-domain planning, the hot start uses the pbest/gbest of the previous round of planning as the initial particle position, and the historical information reuse accelerates convergence by retaining elite particles.

3.2 Emergency rescue path planning based on IGA-IPSO

The path planning for urban emergencies based on IPSO can quickly obtain the shortest and smoothest path for drones to the location of urban emergencies [17]. However, emergency rescue in urban areas not only requires planning the optimal rescue path, but also reasonable task allocation planning for emergency rescue material distribution and other tasks. Multi-drone task allocation: Under the premise of optimal path, to match the best task plan for multiple drones, drone payload and

task urgency are considered [18]. The task allocation problem aims to allocate appropriate tasks to multiple UAVs, with the goal of minimizing the total cost. Constraints include: (1) Load constraint: the total mission load of the UAV is not greater than the maximum mission load; (2) Urgency constraint: the urgency of the task affects the allocation priority; (3) Time window: the task must be completed before the end of the task. These constraints are enforced through the GA encoding and decoding process, or incorporated into the fitness function as penalty terms. However, multi-UAV task allocation technology needs to consider factors such as the load capacity of the UAV and the urgency of the mission target. Multi-drone task allocation technology needs to consider drone payload capacity and the urgency of task objectives. In the multi-UAV task allocation, in addition to load and time window constraints, the research introduces a distributed safety collaboration protocol to ensure that the minimum safe distance between UAVs is always met during task execution. This protocol implements dynamic obstacle avoidance based on the ORCA and combines the communication delay model (OPC UA + MQTT protocol) for real-time collaborative verification. In addition, the task allocation plan needs to pass formal verification tools (such as UPPAAL) for deadlock and conflict detection before execution to ensure its feasibility under high-density dynamic obstacles. The GA can simulate the natural selection and genetic mechanisms of biological evolution. Therefore, the study introduces the GA to assign distribution tasks to each drone in the distribution center, efficiently executing emergency rescue tasks for urban emergencies [19]. The GA framework for multi-drone task allocation is shown in Figure 3.

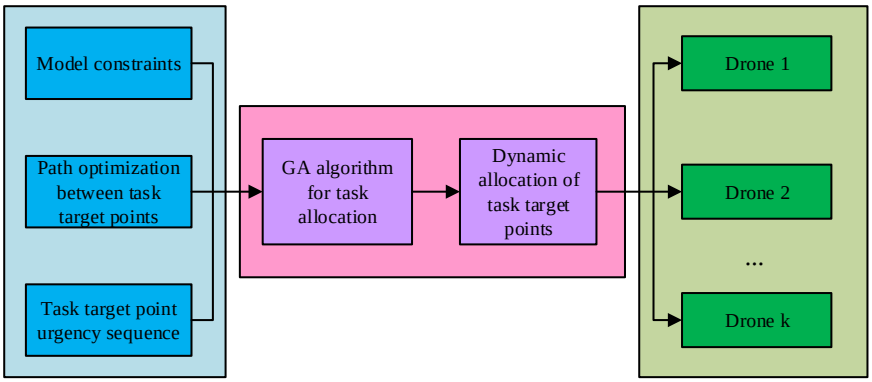


Figure 3: Task allocation structure

In Figure 3, when using the GA to allocate UAV emergency rescue tasks in urban emergencies, it is necessary to first consider model constraints and target point path optimization, and construct emergency sequences to prioritize high-risk points. Then, the GA is used to dynamically allocate tasks and respond to sudden changes, ultimately achieving collaborative execution among multiple drones. Due to the numerous parameters involved in task allocation for multiple drones, GA is difficult to control accurately and has insufficient local optimization performance. Therefore, to optimize the robustness and global search capability of GA during the search process, the GA is optimized. The research introduces a combination strategy of elite retention and roulette to solve the lost excellent solutions in traditional GA during task allocation, and improve the convergence speed and resource allocation efficiency. The expression is shown in equation (9).

$$N_{el} = \lfloor g \times PopulationSize \rfloor \quad (9)$$

In equation (9), N_{el} represents the number of elite individuals. g is the proportionality coefficient. $PopulationSize$ represents the population size. $\lfloor \cdot \rfloor$ is a down rounding function. After the elite retention strategy, the roulette wheel selects the remaining individuals for crossover and mutation operations, and the selection probability of the roulette wheel is shown in equation (10).

$$P(x_i) = \frac{fit(x_i)}{F} \quad (10)$$

In equation (10), $P(\cdot)$ represents the probability of being selected to enter the mating pool. $fit(\cdot)$ signifies the fitness value. F signifies the total fitness of the population. Next, to obtain the cumulative probability distribution interval, the corresponding individuals are selected and the cumulative probability of all individuals being selected by roulette wheel is calculated. The calculation is shown in equation (11).

$$C(x_i) = \sum_{i=1}^{i} P(x_i) \quad (11)$$

In equation (11), $C(\cdot)$ signifies the sum of the selection probabilities of all individuals. $P(\cdot)$ signifies the individual's selection probability. To enhance the local search capability of the GA, the Metropolis criterion is introduced to enhance the local search capability and avoid premature convergence in multi-UAV task allocation. The expression is shown in equation (12).

$$P(\Delta E,T)=\begin{cases}1,\Delta E<0\\ \exp(-\frac{\Delta E}{k_bT}),\Delta E\geq 0\end{cases} \quad (12)$$

In equation (12), P signifies the probability of the algorithm accepting a new solution. ΔE represents the amount of energy change. T represents the current temperature. k_b represents the Boltzmann constant. The GA chromosome uses real number encoding to represent the UAV task allocation sequence. The chromosome length is the number of tasks, and each gene corresponds to a task ID. The crossover operation uses two-point crossover, and the mutation operation is implemented by randomly exchanging genes. Specifically, the crossover probability is 0.8 and the mutation probability is 0.1, which are dynamically adjusted based on the Metropolis criterion. To solve the poor adaptability of fixed search times in complex terrain, the search depth is dynamically adjusted to improve path smoothness and task completion rate. The study introduces an adaptive variable neighborhood search time scheme to deeply optimize the GA. The specific expression is shown in equation (13).

$$IterGen=\rho_1+\left\lfloor \rho_2\times\frac{Gen}{GenMax}\right\rfloor \quad (13)$$

In equation (13), $IterGen$ signifies the total times variable neighborhood searches are performed on elite individuals in the current iteration round. ρ_1 represents the minimum number of guaranteed times for variable neighborhood search. ρ_2 represents the upper limit span of adaptive search times. Gen represents the current iteration count. The GA chromosome uses extended real number encoding and includes fields such as task ID, payload quality, and time window. The reconnaissance task constraint is that the load is no more than 2kg, and the service time is no more than 5min. The material delivery constraint is that the load is no more than 10kg, and the delivery time window is [5,30] min. The types of multi-drone task allocation based on GA are shown in Figure 4.

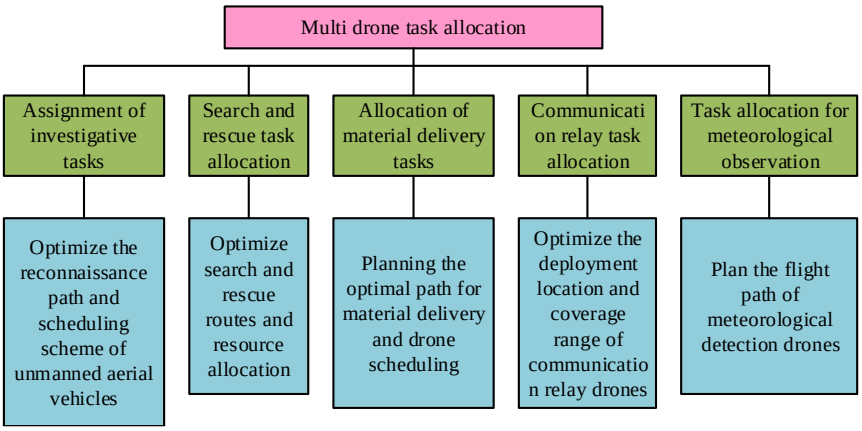


Figure 4: Classification types of urban emergency rescue drone tasks based on GA

As shown in Figure 4, the multi-drone task allocation for urban emergency rescue mainly targets five types of tasks: reconnaissance, search and rescue, material delivery, communication relay, and meteorological detection. GA is used to achieve path planning, resource allocation, scheduling scheme optimization, or deployment location and coverage adjustment. GA is used to improve the scientificity and emergency rescue efficiency of multi-drone task allocation. To determine the urgency sequence of urban emergency rescue needs, the study first introduces a triangular membership function to establish a fuzzy evaluation system, transforming multi-objective conflicts into single-objective optimization, and solving the uneven weight distribution in collaborative decision-making. The expression is shown in equation (14).

$$\mu_{\tilde{a}}(x)=\begin{cases}0,x<a_1 or x>a_3\\ \frac{x-a_1}{a_2-a_1},a_1\leq x\leq a_2\\ \frac{a_3-x}{a_3-a_2},a_2\leq x\leq a_3\end{cases} \quad (14)$$

In equation (14), $\mu_{\tilde{a}}(x)$ represents the degree to which the evaluation object x belongs to the fuzzy concept a. x signifies the actual value of the object to be evaluated. a_1,a_2,a_3 are vertex parameters of the triangle. To further enhance the objectivity of urgency rating, the study clarifies the specific variables and scales of the fuzzy evaluation system.

The set evaluation variables include the number of casualties (unit: people), the impact scope of the event (unit: km^2), and the expected response time (unit: minutes). Each variable is independently rated by experts using a 5-level Likert scale (1-5 points). The parameters of the triangular membership function are calibrated based on historical data: a, b, and c correspond to the minimum, median, and maximum values of each variable score, respectively. To verify the consistency of ratings, the Kappa coefficient is used to evaluate the consistency of ratings among experts, and the ROC curve is used to verify the effectiveness of the ratings in determining actual rescue priority. To determine the urgency weight of emergency rescue in urban emergencies, this study is based on game theory to achieve consistency or compromise between the weights obtained by different weighting methods. The combination weight is presented in equation (15).

$$w = \sum_{k=1}^{n} \alpha_k^* u_k^T \quad (15)$$

In equation (15), w represents the final demand urgency combination weight vector. n signifies the quantity of methods. α_k^* signifies the combination coefficient corresponding to the k-th original weight calculation method. u_k represents the weight vector obtained by the k-th original weight calculation method. This study transforms the path length cost and urgency matching cost of two conflicting objectives into a flexible single objective problem through normalization and weighted summation, providing decision-makers with multiple delivery strategies. Equation (16) presents the objective function.

$$\min F = \theta_1 \frac{f_1 - \min f_1}{\max f_1 - \min f_1} + \theta_2 \frac{\max f_2 - f_2}{\max f_2 - \min f_2} \quad (16)$$

In equation (16), $\min F$ represents the total cost that needs to be minimized. θ_1 and θ_2 represent the importance of balancing the path length cost f_1 with the urgency matching cost f_2, satisfying f_2. Weights θ_1 and θ_2 are dynamically assigned through game theory. The default value θ_1 is 0.6 (path length priority) and θ_2 is 0.4 (urgency priority). Normalization uses Min-Max scaling to ensure that each target value is in the range [0,1], where $\min f_1$ and $\max f_1$ are estimated based on historical data. To eliminate the dimensional difference between the path length cost and the urgency matching cost, the Min-Max normalization is used to scale the two to the [0,1] interval, and its specific expression is shown in equation (17).

$$\begin{cases} C'_{\text{path}} = \dfrac{C_{\text{path}} - C_{\text{path,min}}}{C_{\text{path,max}} - C_{\text{path,min}}} \\ C'_{\text{urgency}} = \dfrac{C_{\text{urgency}} - C_{\text{urgency,min}}}{C_{\text{urgency,max}} - C_{\text{urgency,min}}} \end{cases} \quad (17)$$

In equation (17), C'_{path} and C'_{urgency} are set based on historical extreme scenarios (such as the shortest/longest path, the lowest/highest urgency), and their final objective function is $J = \theta_1 C'_{\text{path}} + \theta_2 C'_{\text{urgency}}$. In summary, the technical roadmap based on the IGA-IPSO is shown in Figure 5.

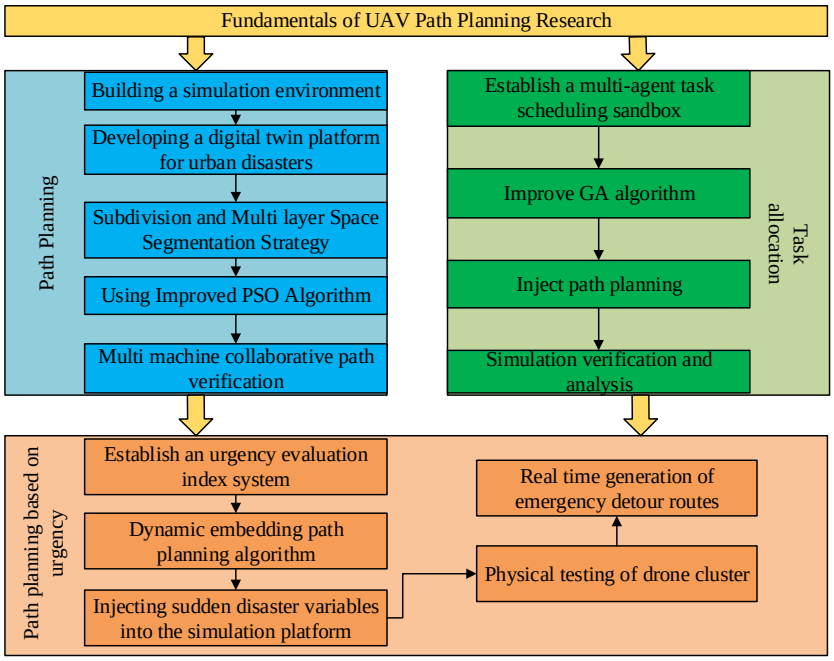


Figure 5: Route planning technology based on improved GA and PSO

In Figure 5, the process based on IGA-IPSO optimizes conventional path planning and multi-machine collaborative verification using hierarchical spatial strategy and IPSO algorithm. In addition, focusing on time critical scenarios, an emergency assessment system is established to coordinate multi-agent task scheduling through IGA, dynamically embed path planning to respond to sudden variables, and achieve "virtual-physical" dual loop verification with drone clusters. The collaboration of two paths enhances the timeliness and adaptability of path planning, providing effective support for emergency rescue in complex urban scenarios. The key parameters involved in the study (including the initial temperature and Boltzmann constant in the Metropolis criterion, as well as the basic number and upper limit span in the adaptive variable neighborhood search) are determined through grid search and pre-experimental calibration. The Metropolis criterion parameters: The initial temperature of 100 is a typical setting based on simulated annealing algorithm combined with exponential cooling strategy. The Boltzmann constant is set to 1 as a standard normalization factor to ensure that the acceptance probability is within a reasonable range. Adaptive variable neighborhood search parameters: Based on sensitivity analysis under different scene complexities (50~200 nodes), the base number is determined to be 5 and the upper limit span is 100 to ensure a balance between convergence speed and search depth. All hyperparameters are tuned through grid search, and the search range is determined based on preliminary experiments. Results are averaged based on 30 independent runs. The hyperparameter settings are as follows:

Table 4: Hyperparameter setting and tuning protocol

Algorithm component	Parameter	Value/Range	Dispatching mode
PSO	Population size	50	Fixed
	Inertia weight	[0.4,0.9]	Adaptive segmentation
	Learning factor c_1, c_2	[0.5,2.0], [0.8,3.0]	Adaptive index
GA	Population size	100	Fixed
	Cross rate	0.8	The Metropolis criteria are dynamically adjusted
	Mutation rate	0.1	The Metropolis criteria are dynamically adjusted
	Selection strategy	Elite reserve + roulette	The proportionality coefficient k=0.1
Variable Neighborhood Search (VNS)	Minimum number of times	5	Adaptive
	Upper limit span	10	Adaptive
Metropolis Criteria	Initial temperature	100	Exponential cooling
Fuzzy evaluation system	Vertex of a triangle (a,b,c)	(0, 0.5, 1)	Calibration based on historical data
Combination coefficients in game theory	λ_i	[0.4,0.6]	Normalized weighting

The specific steps of the IGA-IPSO algorithm are shown in Table 5.

Table 5: Pseudo-code of IGA-IPSO algorithm

Algorithm 1: IGA-IPSO Algorithm
Input: Particle swarm size, maximum number of iterations, fitness function
Output: Optimal path and task allocation
Step 1: Initialization: Generate particle positions using the Tent chaotic map
Step 2: For t = 1 to t_{max}
Step 3: Calculate the fitness
Step 4: Update pbest and gbest
Step 5: If the particle falls into a local optimum, add the traction acceleration
Step 6: Use GA for task allocation
Step 7: If convergence occurs (fitness change <0.01), exit
Step 8: End For
Step 9: Return the optimal solution

As shown in Table 4, the IGA-IPSO algorithm has been described in detail. The research adopts a two-layer loop optimization framework: The outer layer is the IGA task allocation loop, and the inner layer is the IPSO path planning loop. The specific scheduling is as follows: Outer loop (IGA): Each iteration generates a set of task allocation plans. Inner loop (IPSO): Based on the current task plan, calculates the cost of each path. Feedback mechanism: The path cost is fed back to the IGA for updating the fitness function. Stopping criterion: Terminates when the fitness change rate of 10 consecutive rounds of iterations is less than 1% or reaches the maximum number of iterations. This coupling mechanism ensures dynamic coordination between task allocation and path planning.

4 Verification of emergency rescue path planning for urban emergencies based on IGA-IPSO

4.1 Performance testing of emergency rescue path planning based on IGA-IPSO

To verify the performance of the designed method, a simulation model is constructed, and its experimental environment and specific configuration are shown in Table 6.

Table 6: Test environment and specific configuration

Testing environment	Specific configuration
Simulation environment	Python + NumPy
Computing equipment	Intel Core i7-8550U CPU @1.80GHz, 16GB RAM
Environmental monitoring	Distributed gas sensor, and thermal imaging camera
Dynamic obstacle detection	The unmanned aerial vehicle is equipped with a dual-mode visible light/infrared camera
Environmental modeling tool	Three-dimensional topological map
Heterogeneous system integration	The OPC UA + MQTT hybrid protocol is adopted

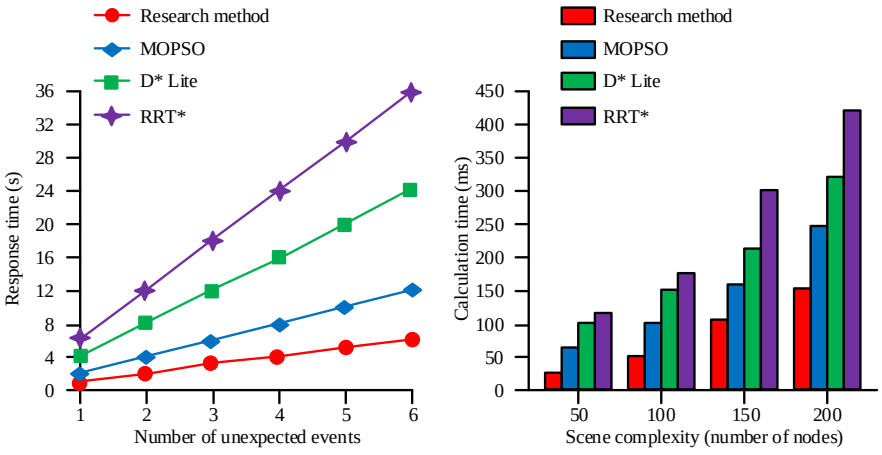
In Table 6, the research uses the experimental environment and specific configuration in the table for performance testing. In this experiment, the performance indicators are clearly defined: The planning success rate is the proportion of the algorithm that successfully generates paths that meet all constraints (such as battery life, no-fly zones, collision avoidance, etc.) in 30 independent runs. Convergence is when the fitness change rate for 10 consecutive iterations is less than 1%, and the algorithm is deemed to have converged. All experimental results are based on 30 independent runs, initialized with different random seeds. Some results are expressed as mean ± standard deviation, and statistical significance is verified by paired t-test (α=0.05). To comprehensively evaluate the robustness of the algorithm in uncertain and dynamic environments, this study injects the following disturbance factors into the simulation: Wind field disturbance: simulated gust model, wind speed 5-10 m/s, random direction; Positioning noise: injected Gaussian noise (σ=0.5m) in position perception; Dynamic obstacles: used random walk model, speed 0-5 m/s, acceleration ±1m/s²; Communication delay: simulated OPC UA + Communication delay under MQTT protocol (average 100ms, standard deviation 20ms); Map error: randomly offset obstacle position (offset obeys N(0,1)m). Evaluation indicators include: Feasibility rate: the proportion of successfully generating feasible paths in 30 runs; Number of safety constraint violations: such as collisions, no-fly zone intrusions, etc.; Mission deadline miss rate: the proportion of task completion time exceeding the preset time window. To ensure the repeatability of the experiment, the study uses the public road network data set (OpenStreetMap, a city center area, area 10km×10km) as the basic map. The map grid resolution is 100m×100m, the elevation data comes from SRTM DEM (spatial resolution 30m), and the no-fly layer (such as high-rise buildings, sensitive areas) is synthesized. The dynamic obstacle generation is based on the random walk model, the density unit is "person/km²", the initial position of the obstacles is randomly distributed, the movement speed range is 0-5m/s, and the update frequency is 1Hz. Scenario complexity is expressed as the number of nodes (50-200) in the path planning graph, where nodes represent road intersections or task points. Based on map density, number of obstacles and task complexity, the study defined three difficulty levels. The details are shown in Table 7.

Table 7: Definition of difficulty levels

Difficulty level	Map type	Obstacle density (units /km²)	Number of tasks	The number of risk areas
Tier 1	Low-density area	20	1-2	0-1
Tier 2	Low-density area	40	3-4	2-3
Tier 3	High-density area	60	5-6	4-5

In Table 7, 30 independent experiments were run at each difficulty level, and indicators such as planning success rate, response time, and path length were recorded. The scene generation and parameter setting in the test mainly include: (1) Concurrent event generation: Randomly generate 1-6 emergency event points in the map. Event types include fire, car accident, drowning, etc., and the event locations obey uniform distribution; (2) Dynamic obstacle density: Simulate urban dynamic traffic by controlling the number of obstacles per unit area (such as 40/km²), and the obstacle motion model is random direction + constant speed; (3) Scene complexity: Control the scale of the problem by adjusting the number of nodes in the path planning diagram (50, 100, 150, and 200), where nodes represent accessible path points; (4) Risk map synthesis: Construct a risk weight map based on elevation data, obstacle density and historical event data, and calculate the risk value; (5) Task urgency and time window: Task urgency is generated based on the triangle membership function, and the time window is randomly assigned through a uniform distribution (5-30 minutes). The research method is compared with the

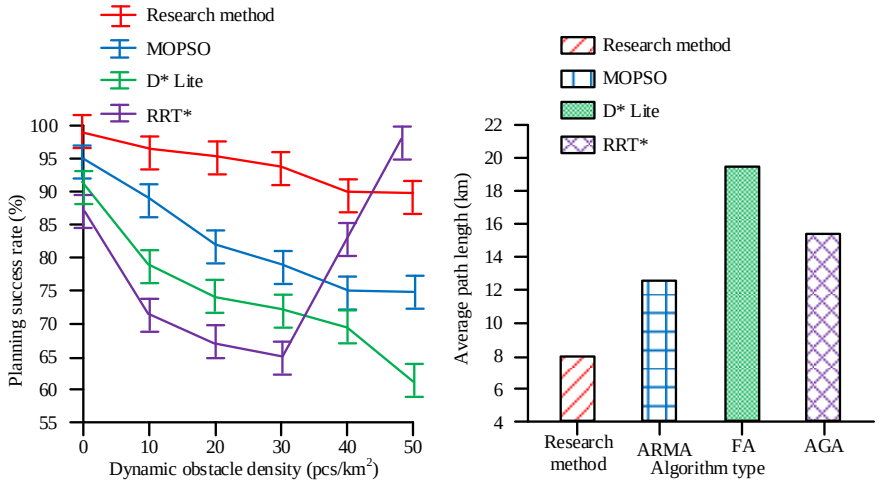
Rapid Exploration of Random Tree Stars (RRT*), D* Lite and the Multi-objective Particle Swarm Optimization Algorithm (MOPSO). The RRT* algorithm is a sampling-based path planning algorithm, suitable for high-dimensional space and dynamic environments. The D* Lite algorithm is an incremental search algorithm, suitable for dynamic obstacle environments. The MOPSO algorithm is a representative of multi-objective optimization algorithms, used to compare the performance of multi-UAV task allocation. All baseline methods use the same cost function (path length and risk-weighted sum) and constraints (drone endurance, obstacle density, etc.) for fair comparison. To ensure the fairness of the experiment, the maximum running time budget is set to 300 seconds per task. If the algorithm does not converge within the budgeted time, it is considered a failure. All algorithms are executed in single-threaded mode to ensure fair comparison. The experimental platform does not have multi-threading or GPU acceleration enabled. The response time of the four methods under different numbers of emergencies and the single planning calculation time under different scene complexities are compared, as presented in Figure 6.



(a) Dynamic programming response time (b) Single planning calculation time

Figure 6: Dynamic programming response time and single planning calculation time

In Figure 6 (a), the response time of path planning for all four methods gradually increased with the increase of unexpected events, with the research method showing the slowest increasing trend. When the number of emergencies was 1, the response time of the research method was 1 second. When the emergencies increased to 6, the response time increased linearly from 1 to 6 seconds. However, the response time of the three comparison methods was higher than. In Figure 6 (b), the single planning calculation time of the four methods varied under different scene complexities. When the scene complexity was 50 nodes, the computation time of the research method was 25±2ms, and the computation time of the three comparison methods was not obviously different from that of the research method. When the scene complexity was 200 nodes, the single computation time of the research method was 150±12ms, while the computation time of the other three methods was much longer. Overall, compared to comparison methods, the research method has better stability and planning efficiency. The success rates of four methods in planning under different dynamic obstacle densities, as well as the average path lengths planned by the four methods, are compared, as presented in Figure 7.



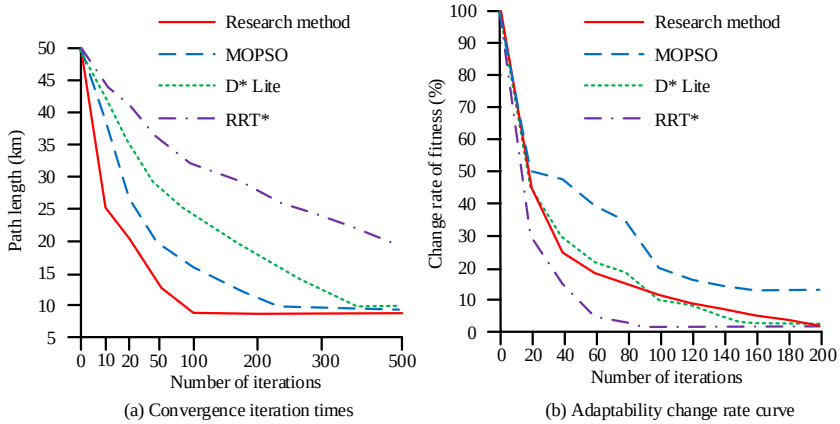
(a) Algorithm success rate (b) Average path length

Figure 7: Algorithm success rate and average path length

In Figure 7 (a), the planning success rates of the four methods decrease as the density of dynamic obstacles increases. In 30 independent runs, when the dynamic obstacle density was 40/km², the planning success rate of the research method (that is, the proportion of feasible paths successfully generated) reached 90.2%. However, the planning success rates of the other three methods were significantly lower, with the RRT* showing an abnormal increase. In Figure 7 (b), the obvious difference existed in the average path length of the rescue routes planned by the four methods. The average path length of the research method was 8.0±1.2km. Among the three comparison methods, the MOPSO model had an average path length of 12.5±1.5km, while the D* Lite and RRT* had average path lengths of 19.7±2.2km and 15.4±1.6km, respectively, which were significantly larger than that of the research method. Overall, compared to comparison methods, the research method has stronger robustness and optimization capabilities. To verify the impact of constraints on algorithm performance, the study tested the planning success rate under different battery life (from 10km to 50km) and no-fly zone density (from 5/km² to 30/km²). The results show that when the battery life is less than 20km, the planning success rate drops to 70%. When the no-fly zone density exceeds 20/km², the path length increases by an average of 15%. This highlights the importance of constraint handling in complex scenarios. All experimental results are based on 30 independent runs, initialized with different random seeds. Some results are expressed as mean ± standard deviation, and statistical significance is verified by paired t-test (α=0.05). The details of instance generation are: (1) Emergency event types include fires, car accidents, drowning, etc., and the locations are uniformly distributed; (2) Dynamic obstacles are generated based on the random walk model, with a movement speed of 0-5m/s, and an update frequency of 1Hz; (3) The complexity of the scene is controlled by the number of nodes (50-200), and the nodes represent passable path points; (4) The risk map is synthesized based on elevation, obstacle density, and historical event data. Moreover, to verify the robustness of the algorithm under extreme conditions, safety-critical stress test scenarios are designed: (1) Sensing noise: Injecting Gaussian noise (σ=0.5m) into position sensing; (2) Execution noise: Adding random deviation (±5%) to the control instructions; (3) Wind field disturbance: Simulating gust model (wind speed 5-10m/s, random direction). The test results show that under the dynamic obstacle density of 40/km² and wind field disturbance, the planning success rate of the research method is 87.5%, which is only 2.7% lower than that of the non-disturbance scenario. The success rate of the comparison methods (RRT*, D* Lite) dropped by more than 15%. In addition, after introducing CBF and ORCA, the number of collisions dropped from an average of 1.2 times per task to 0.1 times, verifying the effectiveness of the safety filter. The IGA-IPSO has the best planning efficiency, stability, and robustness.

4.2 The practical application effect of emergency rescue path planning based on IGA-IPSO

On the basis of verifying the performance of the IGA-IPSO, the practical application value is verified. The research adopts the urban operation basic dataset and builds a multi-modal simulation platform. The research method is compared with MOPSO, D* Lite, and RRT*. The path length and fitness value change rate of the four methods are compared at different iteration times, and the results are shown in Figure 8.

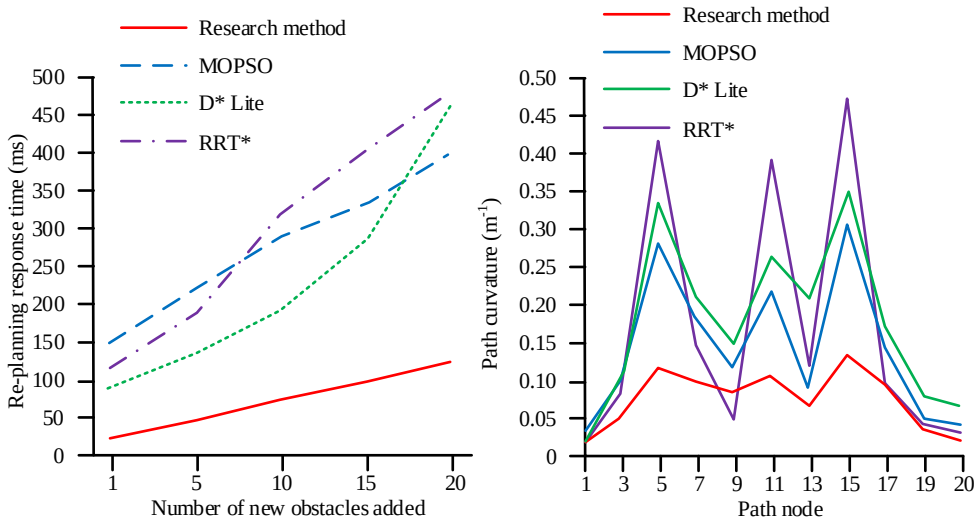


(a) Convergence iteration times (b) Adaptability change rate curve

Figure 8: Curve of convergence iteration times and fitness change rate

In Figure 8 (a), the path lengths of the four methods gradually converge as the number of iterations increases. According to the convergence criterion (fitness change rate <1%), the research method completed convergence at the 100th iteration, and the path length reached a stable value of 8km. However, the convergence speed of the other three methods was significantly slower. In Figure 8 (b), the fitness change rate of the four methods gradually decreased with the increase of iteration times. The fitness change rate of the research method decreased gently, reaching 25.1% at the 40th iteration and 1.3% at the 200th iteration, with stable convergence. The other three

methods showed poor convergence, quadratic convergence, and fast convergence, respectively. Overall, compared to comparison methods, the research method has better efficiency and optimization capabilities. The response time and path curvature changes of four methods for re-planning are compared, as displayed in Figure 9.



(a) Re-planning response time (b) Path smoothness

Figure 10: Response time and path smoothness for re-planning

In Figure 9 (a), the re-planning time of the four methods increased with the increase of the number of new obstacles added. The research method had a re-planning response time of 25±2ms when the obstacle was added once, and 125±10ms when the obstacle was added 20 times, with an increase of 100±8ms. The re-planning time of the three comparison methods was significantly greater. In Figure 9 (b), the four methods had different curvatures at different path nodes. The research method had a maximum curvature of 0.13±0.02m^{-1} at node 15, and peaked in curvature at nodes 5 and 11. The overall curvature of the other three methods at 20 nodes was greater. Moreover, in an actual environment with wind fields and sensor noise, the average re-planning time of the research method was 135ms (obstacles are added 20 times), which was only an 8% increase compared to the undisturbed scene. The path curvature remained within 0.15m^{-1} under wind field disturbance, indicating that the path smoothness was not significantly affected. In addition, through CBF real-time correction, the drone successfully avoided all dynamic obstacles and no-fly zone intrusions during flight. Overall, compared to comparison methods, the research method has better reliability and stability. All experimental results are based on 30 independent runs, initialized with different random seeds. Some results are expressed as mean ± standard deviation, and statistical significance is verified by paired t-test (α=0.05). Finally, to evaluate the independent contribution of each improvement component, ablation experiments were designed. The experiment was conducted under the same instance (six concurrent events, dynamic obstacle density 40/km^2) and fixed random seeds, and the following components were gradually added: Tent chaotic mapping (C), adaptive segmented inertia weight (W), exponential learning factor (L), traction acceleration (T), elite retention and roulette (E), Metropolis criterion (M), and adaptive variable neighborhood search (V). The results show that traction acceleration (T) and adaptive variable neighborhood search (V) contribute the most to avoiding local optimality, while chaotic initialization (C) and elite strategy (E) significantly improve the convergence speed. The IGA-IPSO has the best efficiency, performance, reliability, and stability. To verify the impact of weights θ_1 and θ_2 on the planning results, a sensitivity experiment was designed: fix $\theta_1+\theta_2=1$, adjust θ_1 from 0.0 to 1.0 with a step size of 0.1, and evaluate the changes in path length, urgency matching error and total cost. The results show that: when $\theta_1>0.7$, path length takes priority, but the urgency matching error increases by more than 15%. When $\theta_1<0.3$, urgency matching takes priority, but the path length increases by more than 20%. When $\theta_1\in[0.4,0.6]$, the total cost changes slowly (fluctuation <5%), verifying the robustness of the default weight. This analysis provides decision-makers with a basis for weight tuning to adapt to the preferences of different rescue scenarios.

5 Discussions

The research proposes an urban emergency rescue path planning method based on the improved GA and the PSO algorithm (IGA-IPSO), and its performance is verified through simulation experiments. (1) Performance comparison and attribution with existing technology: The experimental results that the average path length of this research method is 8.0 km, while the

average path length of the comparison method is significantly longer. This improvement is mainly attributed to several optimization components introduced in the IGA-IPSO algorithm: (i) Tent chaotic mapping. (ii) Traction acceleration term. (iii) Adaptive piecewise inertia weights and exponential learning factors. (iv) Elite Retention and Metropolis Code. (2) Trade-offs and limitations: Although the present research method performs well in path length, convergence speed, and robustness, there are still trade-offs and limitations to consider: (i) Computational complexity. (ii) Parameter sensitivity. (iii) Robustness. (3) Scalability analysis. The research tests up to six concurrent events and dynamic obstacle densities up to 40/km² in experiments. From the results: (i) Event number expansion: Response time increases linearly as the number of events increases. (ii) Map density expansion: When the scene complexity is up to 200 nodes, the research method still maintains low calculation time and high success rate. (iii) Multi-UAV collaboration: Through distributed security protocols and formal verification, the research method supports multi-UAV task allocation and path planning. (4) Summary and outlook: In summary, this research method is superior to the existing technology in path length, convergence speed and robustness by integrating improved PSO and GA components. Its main advantage is that it dynamically balances multi-objective optimization and constraint processing, and is suitable for complex urban emergency rescue scenarios. Future work will focus on path planning under extreme weather, large-scale event expansion, and real-time learning mechanisms.

6 Reproducibility and experimental materials

To ensure the repeatability and transparency of this research experiment, complete code, data, configuration and experimental materials are provided and publicly available in the GitHub repository (repository address: https://github.com/[username]/IGA-IPSO-Emergency-Path-Planning). The repository contains the following content:

(i) Deterministic Seeds: All experiments use fixed random seeds (such as seed=42) to ensure consistent results for each run. (ii) Scenario Generators and Maps provide road network data (in .osm or .graphml format) based on OpenStreetMap, and the dynamic obstacle generation script scenario_generator.py. The script supports custom event points, obstacle density and motion models. (iii) Configuration Files (Configuration Files): All algorithm hyperparameters, environment configurations and experimental parameters are stored in the config/ directory in YAML format, including PSO algorithm parameters (such as inertia weights, learning factors, etc.), GA parameters (such as crossover rate, mutation rate, etc.), and simulation environment configuration (such as map range, resolution, obstacle attributes, etc.). (iv) Reproduction Scripts provides a complete Python script for reproducing all charts and tables: The main experimental script performs performance testing and practical application verification, generates all result graphs, and outputs tabular data. (v) README file: README.md in the root directory of the warehouse contains detailed environment configuration instructions, including Python 3.8+ environment setup, dependent library list (such as NumPy, Matplotlib, NetworkX, etc.), data preprocessing steps, and step-by-step operation guide. All materials have been tested and can reproduce all experimental results in this article under the same hardware and software environment.

7 Conclusion

In response to the frequent occurrence of urban emergencies, the low efficiency of traditional emergency rescue path planning, and the insufficient ability to coordinate multiple tasks, an innovative urban emergency rescue path planning method based on the IGA-IPSO was proposed. Compared with traditional Tent chaotic mapping, this study uses parameter adaptation and scene calibration mechanisms to make the initialization process more in line with the dynamic needs of urban emergency rescue, effectively improving the convergence speed and robustness of the algorithm. The research method incorporated the traction acceleration term to optimize the PSO, designed an adaptive variable neighborhood search mechanism to optimize the GA, and constructed a fuzzy evaluation system for urgency. The matching cost between path length and urgency was transformed into a single objective optimization problem. When the scene complexity was 50 nodes, the computation time was 25±2ms. The average path length was 8.0km. In practical application testing, the fitness change rate of the research method decreased gently, from 25.1% in the 40th iteration to 1.3% in the 200th iteration, with stable convergence. The research method showed that the rescue cost for car accidents was around CNY 2,000±12. The rescue cost during a drowning incident was around CNY 500±8. The rescue costs for fire and disease incidents were around CNY 3,500±20 and CNY 2,500±15, respectively. When an obstacle was added once, the response time for re-planning was 25±2ms. When an obstacle was added 20 times, the response time for re-planning was 125±10ms, with an increase of 100±8ms. The research significantly improves the robustness of path planning in safety-critical scenarios by introducing the Control Barrier Function (CBF), ORCA obstacle avoidance algorithm and wind field disturbance model. The stress test shows that the research method still maintains a high success rate and low collision rate under sensing noise, execution deviation and wind field interference, which is suitable for complex dynamic environments in actual urban emergency rescue. Overall, the proposed method enhances the timeliness, robustness, and economy of emergency rescue in complex urban conditions. The research method significantly improves the feasibility and safety of path planning by formalizing key constraints such as battery life, no-fly zones, and collision avoidance. Experiments show that the constraint

processing mechanism still maintains a planning success rate of more than 90% under dynamic obstacles and high-density no-fly zones. The current research does not consider the interference of extreme weather on the UAV dynamic model. In future research, Reinforcement Learning (RL) can be introduced for dynamic policy learning and environmental adaptive decision-making to further improve the robustness of the path planning system in uncertain environments. Specifically, RL can be used to: (1) Dynamically adjust the inertia weight and learning factor in IPSO so that it changes adaptively according to real-time wind field, visibility and other meteorological conditions; (2) Optimize the task allocation strategy in IGA and dynamically adjust task priority and resource scheduling according to the evolution of real-time events; (3) Collaborate with the existing IGA-IPSO framework to form a two-layer decision-making mechanism of "offline optimization + online learning": IGA-IPSO is responsible for global path and mission planning, and RL is responsible for local dynamic adjustment and disturbance compensation. By embedding RL into the existing framework and combining it with the meteorological-terrain coupling constraint model, the research method can better adapt to dynamic uncertain scenarios such as extreme weather, and improve the real-time performance and adaptability of the emergency rescue system.

8 Deployment considerations

The main deployments in the study include: (1) Regulatory compliance. BVLOS (beyond visual line-of-sight) approval: Apply for a BVLOS flight permit from the Civil Aviation Administration in actual deployment, and submit a risk assessment report, emergency response plan, and other materials; Dynamic update of no-fly zones: The system should access real-time no-fly zone data released by civil aviation and the military, and respond dynamically in path planning; Airspace hierarchical management: Design an airspace priority usage mechanism based on the urgency of the task to ensure priority for high-urgency tasks. (2) Safety interlocking mechanism. Real-time health monitoring: The drone needs to be equipped with battery voltage, GPS signal, communication link and other status monitoring. Once abnormal, it will trigger automatic return or backup; Human intervention interface: The system should support manual takeover by the command center, with functions such as one-click pause, path re-planning, and mission abort; Fault emergency protocol: Define the emergency handling process for extreme situations such as the drone losing contact, crashing, and being hijacked. (3) System integration and command and dispatch. API interface specifications: Provide RESTful API or MQTT themes to support docking with urban emergency command platforms (such as 110, 120, and 119 systems); Task status synchronization: Push task progress, path deviations, remaining resources and other information to the command center in real time; Logs and auditing: Record all planning decisions, path execution, abnormal events, and support post-event review and responsibility tracing. (4) Multi-modal collaboration and handover protocol. UAV-ground vehicle collaboration: (i) UAVs are responsible for rapid reconnaissance and preliminary material delivery, and ground vehicles are responsible for large-scale material transportation or personnel transfers; (ii) Handover points are usually located in safe areas (such as squares and school playgrounds), and UAVs guide ground vehicles to the target point. Mission handover process: (i) After the drone completes the mission, it sends "mission completion + location information" to the command center; (ii) The command center assigns the nearest ground vehicle to respond; (iii) The drone hovers or lands at the handover point, waiting for confirmation before returning.

Funding

The research is supported by Special Project of Guangxi Public Security Department in2023: Research on Prevention and Control Strategies for the New Business Model of Homestay +"Accommodation in Guangxi Border Areas (2023GAYB048); 2024 Guangxi Higher Education Undergraduate Teaching Reform Project: "Student centered" New Era Public Security College General Education Upgrade Path, Reform and Practice of Dual wheel Driven Curriculum System(2024JGB426); 2024 Guangxi Higher Education Undergraduate Teaching Reform Project: Teaching Reform and Practice of "Specialized Creation Integration" in the Course of Public Management under the Background of New Liberal Arts(2024JGA400).

References

[1] Xu W, Xu J, Proverbs D, Zhang Y. A hybrid decision-making approach for locating rescue materials storage points under public emergencies. Kybernetes, 2024, 53(1): 293-313. DOI:10.1108/k-07-2022-1060

[2] Rhodes C, Liu C, Westoby P, Chen W H. Autonomous search of an airborne release in urban environments using informed tree planning. Autonomous Robots, 2023, 47(1): 1-18. DOI:10.1007/s10514-022-10063-8

[3] Peixoto J P J, Costa D G, Portugal P, Vasques F. Flood-resilient smart cities: a data-driven risk assessment approach based on geographical risks and emergency response infrastructure. Smart Cities, 2024, 7(1): 662-679. DOI:10.3390/smartcities7010027

[4] Sun B, Zhou Y. Bayesian network structure learning with improved genetic algorithm. International Journal of Intelligent Systems, 2022, 37(9): 6023-6047. DOI:10.1002/int.22833

[5] Yun Y, Gen M, Erdene T N. Applying GA-PSO-TLBO approach to engineering optimization problems. Math. Biosci. Eng, 2023, 20(1): 552-571. DOI:10.3934/mbe.2023025

[6] Jose C, Vijula Grace K S. Optimization based

routing model for the dynamic path planning of emergency vehicles. Evolutionary Intelligence, 2022, 15(2): 1425-1439. DOI:10.1007/s12065-020-00448-y

[7] Zhao X, Wang G. Deep Q networks-based optimization of emergency resource scheduling for urban public health events. Neural Computing and Applications, 2023, 35(12): 8823-8832. DOI:10.1007/s00521-022-07696-2

[8] Liu J, Li Y, Li Y, Zibo C, Lian X, Zhang Y. Location optimization of emergency medical facilities for public health emergencies in megacities based on genetic algorithm. Engineering, construction and architectural management, 2022, 30(8): 3330-3356. DOI:10.1108/ecam-07-2021-0637

[9] Xia H, Sun Z, Wang Y, Zhang J Z, Kamal M M, Jasimuddin S M, et al. Emergency medical supplies scheduling during public health emergencies: algorithm design based on AI techniques. International Journal of Production Research, 2025, 63(2): 628-650. DOI:10.1080/00207543.2023.2267680

[10] El-Shafiey M G, Hagag A, El-Dahshan E S A, Ismail M A. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. Multimedia Tools and Applications, 2022, 81(13): 18155-18179. DOI:10.1007/s11042-022-12425-x

[11] Shen L, Zhang Z, Song X, Yin Y. A hybrid GA-PSO algorithm for seru scheduling problem with dynamic resource allocation. International Journal of Manufacturing Research, 2023, 18(1): 100-124. DOI:10.1504/ijmr.2023.10040833

[12] Bousnina K, Hamza A, Ben Yahia N. A combination of PSO-ANN hybrid algorithm and genetic algorithm to optimize technological parameters during milling 2017A alloy. Journal of Industrial and Production Engineering, 2023, 40(7): 554-571. DOI:10.1080/21681015.2023.2243312

[13] Sun H, Li H, Meng Z, Wang D. An improvement of dv-hop localization algorithm based on improved adaptive genetic algorithm for wireless sensor networks. Wireless Personal Communications, 2023, 130(3): 2149-2173. DOI:10.1007/s11277-023-10376-6

[14] Zhi Y, Wang H, Wang L. A state of health estimation method for electric vehicle Li-ion batteries using GA-PSO-SVR. Complex & Intelligent Systems, 2022, 8(3): 2167-2182. DOI:10.1007/s40747-021-00639-9

[15] Aloqaily M, Elayan H, Guizani M. C-healthier: A cooperative health intelligent emergency response system for c-its. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(2): 2111-2121. DOI:10.1109/tits.2022.3141018

[16] Pillai B M, Suthakorn J, Sivaraman D, Nakdhamabhorn S, Nillahoot N, Ongwattanakul S, et al. A heterogeneous robots collaboration for safety, security, and rescue robotics: e-ASIA joint research program for disaster risk and reduction management. Advanced Robotics, 2024, 38(3): 129-151. DOI:10.1080/01691864.2024.2309622

[17] KA A, Subramaniam U. A systematic literature review on multi-robot task allocation[J]. ACM Computing Surveys, 2024, 57(3): 1-28. DOI:10.1145/3700591

[18] Abujabal N, Fareh R, Sinan S, Baziyad M, Bettayeb M. A comprehensive review of the latest path planning developments for multi-robot formation systems[J]. Robotica, 2023, 41(7): 2079-2104. DOI:10.1017/s0263574723000322

[19] Singh G, Chaturvedi A K. Hybrid modified particle swarm optimization with genetic algorithm (GA) based workflow scheduling in cloud-fog environment for multi-objective optimization. Cluster Computing, 2024, 27(2): 1947-1964. DOI:10.1007/s10586-023-04071-1

A Proximal Policy Optimization-Based Reinforcement Learning Framework for Real-Time Personalized Endurance Training

Chuanzhong Wu[1], Danqing Liang[2], Bo Yang[2], Li Xu [1], Yunlong Li[2,*]
[1]Guangzhou College of Commerce, Guangzhou, Guangdong,202162, China
[2]Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai ,202162, China
E-mail: liyunlong998@hotmail.com
[*]Corresponding author

Keywords: reinforcement learning, personalized training, sports optimization, wearable sensors, athletic performance

Recieved: July 11, 2025

Customized sports training routines take into account individual physiology, fatigue, and recovery to maximize performance. Proximal Policy Optimization (PPO)-based reinforcement learning is used to adjust training intensity, duration, and rest in a simulated endurance-training environment for runners, using real-time wearable and performance data. The environment models athlete status utilizing heart rate variability, VO_2 max, fatigue ratings, and injury-risk indicators. PPO is trained to maximize performance gains, recovery quality, and safety over repeated sessions. Simulated policy improves performance (18.6%), injury-risk deviation (−22.4%), recovery compliance (91.3%), training load variability control (±7.2%), reward-signal evolution convergence (+41.7%), session completion rate (94.6%), personalized adaptation score (87.5%), and fatigue index stability (94.3%). Results show that a PPO-based RL setup, specifically defined by state design, reward shaping, and multi-episode training, can provide adaptive and data-driven tailored sports training.

Povzetek: PPO-temelječ pristop z okrepljenim učenjem na podlagi podatkov nosljivih naprav dinamično prilagaja tekaški trening (intenzivnost, trajanje, počitek) ter v simulacijah hkrati izboljša zmogljivost, okrevanje in varnost.

1 Introduction

Optimizing athletic performance requires adaptive training methods that account for individual variations in physiology, recovery, and mental preparedness. A conventional static program is not responsive to time-sensitive changes in an athlete's condition [1]. This paper introduces a new reinforcement learning strategy, based on PPO, that personalizes all training parameters using wearable biometric data. The system is designed to facilitate performance, enable faster recovery, and minimize the risk of injury by dynamically varying intensity, duration, and rest, introducing an intelligent, data-driven dimension to sports training [2].

1.1 Background

Optimizing athletic performance has long been a source of research, driving the development of new training techniques, including periodized conditioning and sport-specific training [3]. Conventionally, coaches and athletes have relied on pre-stipulated schedules founded on experience, general physiological principles, and observations. Although, to a certain extent, these methods can be described as successful, these procedures do not consider the individuality and changing nature of human physiology, individual recovery, or even the stress response. The need to tailor training to the individual is becoming an urgent necessity in the contemporary high-performance context [4].

1.2 Motivation

Elite and amateur athletes alike are constantly striving to deliver their best performance and minimize the risks of declining performance, injury, and burnout. Even the most convenient static training programs do not adjust in real time to key factors such as fatigue, sleep quality, heart rate variability, and mental preparedness. The need for such a study stems from the growing recognition that adaptive systems based on artificial intelligence can address this gap by dynamically adjusting training protocols in real time based on ongoing feedback [5].

1.3 Importance of personalization in sports training

Athletes cannot be made equal. Responding to a training stimulus depends on various factors, including age, genetics, injury history, training history, and psychological factors [6]. Personalization ensures that every sports person receives the optimal training, intensity, and volume tailored to their body and ambitions. Individualized programs are not only faster in terms of gains, but also more compliant, reduce the incidence of injuries, and promote long-term growth. Due to the widespread

adoption of wearable devices and sensor technologies, it is now possible to design training interventions guided by granular and individualized data [7].

1.4 The rise of data-driven athletic interventions

Sports activities have led to an influx of biometric data captured by smartwatches, GPS trackers, and heart rate monitors, among other devices. They are collecting constant streams of data: everything from steps and sleep stages to lactate threshold and recovery percentage. The new flow of real-time data creates opportunities for data-driven training management, especially when combined with more advanced algorithms, such as machine learning and reinforcement learning. Optimally integrated, this data can be used to perform dynamic interventions, rather than applying the given rules [8].

1.5 Limitations of existing training methods

The training methods used now are primarily based on a few heuristics or previous trends and thus do not account for the immediate physiological effects. These procedures do not account for day-to-day variations in the athlete's readiness; therefore, load management is poor [9]. Additionally, rule-based or conventional machine learning systems may require extensive manual tuning and may struggle to keep pace with nonlinear or long-term feedback trends. These systems can be optimized to work immediately but fail to provide adequate progression, rehabilitation, and injury prevention [10].

1.6 Research contributions

- To create a real-time PPO-based reinforcement learning model that adjusts training intensity, duration, and recovery depending on heart rate variability, VO_2 max, and fatigue, resulting in a tailored and responsive training regimen.
- To assess if the PPO framework achieves a minimum 18.6% performance increase (VO_2 max proxy) over static or rule-based endurance training regimens.
- To determine if the PPO-driven method can minimize injury risk deviation by 22.4% and increase session completion and recovery compliance to 94.6% and 91.3%, exceeding DQN, A3C, JITAI, and other RL baselines.

To assess if integrating a weighted reward function with PPO clipping leads to stable load variability (±7.2%), strong individualized adaptation scores (87.5%), and high fatigue index stability (94.3%) with increasing sample size.

1.7 Paper organization

In this paper, section 2 explains the related works, section 3 provides the methodology, section 4 shows the evaluation metrics. Result analysis is explained in section 5, and the discussion is shown in section 6. Finally, section 7 suggested the conclusion with future works.

2.Related works

In recent years, the application of artificial intelligence, particularly reinforcement learning and deep neural networks, has increased in sports training. Key techniques include Deep Q-Networks and Asynchronous Advantage Actor-Critic (A3C), applied across various levels of athlete performance. While these studies show promise in intelligent adaptation, most lack real-time sensor feedback, closed-loop systems, or personalized training variables. This literature review examines 10 recent techniques, highlighting their strengths, limitations, and significant gaps that necessitate a dynamic framework grounded in PPO.

In this research, a Deep Reinforcement Learning (DRL) algorithm with an Asynchronous Advantage Actor-Critic (A3C) framework will be designed to learn and tune exercise performance. The system speeds up or reduces the extent of training based on a trade-off between exploration and exploitation, and may learn to model fitness-fatigue dynamics by analyzing time-series biometric data. Validation was done in publicly available walking, running, and sports logs. The power of DRL to individualize training plans stems from its ability to transform exercise prescriptions to reflect users' real-time fitness levels [11].

The authors implement the Deep Q-Network (DQN) to personalize the teaching path in physical education using data from students' activity logs. Once a data set is gathered and processed, the model establishes a system of definite mappings between state-action-reward associations, enabling dynamic adjustment of PE training. The DQN approach yields significantly better autonomy, participation, and skill results than traditional static teaching approaches. The experiment has demonstrated that customized training programs, developed with the aid of RL, are superior to generic curricula in learning physical education [12].

The paper proposes Expected Discounted Goal (EDG), a deep reinforcement learning method for assessing the contributions of soccer players in simulation settings. Without any real-life training data in the form of human labels, the model relies solely on virtual match simulations and open-source tracking. EDG estimates the number of goals a team may score or concede per state, which helps to analyze player performance more effectively. The presented strategy can be applied in other sports where labeled training data are scarce, as noted in Garnier et al [13].

This was achieved by developing a wearable AI system that provides real-time biomechanical feedback during hammer-throwing training. It combines inertial sensors, load cells, and deep neural networks (DNN) to predict limb position angles and wire tension. For this system, an objective evaluation of motion is provided, with subjective feedback based on experience replaced by an objective correction based on the data. With feasible estimation inaccuracies of under 12%, the paper demonstrates how DNNs and wearables can enable

dynamic training on complex athletic movements tailored to individual needs [14].

This paper examines AI-driven advancements in sports science, concentrating on load management, injury prevention, talent identification, and recovery efficiency. It supports application-specific artificial intelligence systems that augment human knowledge and expertise. They can be used to predict overtraining, the menstrual cycle, and mental health. The research also recommends interdisciplinary cooperation and revised educational programs to make future sport scientists' literate about AI. It highlights the importance of transparency and the ethical application of AI in high-performance environments [15].

The paper proposes an algorithm for reinforcement learning of Just-In-Time Adaptive Intervention (JITAI) via the mobile app HeartSteps V2. The RL model personalizes physical activity recommendations and provides them as often as 5 times a day, depending on real-time user context. Continuous learning in health applications is effective because the system learns and improves over time as more user behavior data is received. The JITAI method can also be utilized in dynamic training plans for sports or fitness [16].

This paper combines Q-learning with upper-limb training (a VR bubble-popping game). Each session is adaptive and patient-specific because difficulty is adjusted based on the patient's actions, using kinematic information. The algorithm optimised bubble locations and timing based on patient capability over 10 sessions, indicating the potential for intelligent physical therapy. It is a model of reinforcement learning to assist in motor recovery and personalized feedback in exercise therapy [17].

The paper applies Subgroup Discovery (SD) to identify patterns between training load, wellness parameters, and overuse injuries among a group of professional volleyball players. The model determines the personal risks of injury by measuring wellbeing and workload data of athletes who are followed for 24 weeks. Individualized signs, such as sleeping less or exhibiting a jump load behavior, were noted. That is the value of frequent observation and flexible planning, which can be further improved through reinforcement learning to make dynamic adjustments [18].

This paper utilizes Convolutional Neural Networks (CNNs) and Transfer Learning to categorize complex throwing motions in Ultimate Frisbee using IMU data. It achieves an accuracy of up to 89.9 percent in inferring the three main types of throws and benefits from pre-set weights when the data is scarce. Although it is not RL-based, the ability to identify fine-grained athletic actions enables proper state recognition in RL models, underscoring the role of CNNs in performance monitoring [19].

2.1 Research gap

Limitations of the related works is shown in Table 1.

Table 1: Limitations of the related works

S.No.	Paper / Method	Key idea & quantitative result (where given)	Main limitation	Gap vs. this work (no closed-loop sensor RL)
1	DRL fatigue–recovery (A3C)	RL co-models' fatigue and recovery; improves simulated reward and reduces over-fatigue episodes vs. heuristics.	Simulation only; uses abstract states, no rich wearable input.	Does not run a real-time loop from sensor state → training action → updated state for full-plan personalization.
2	DQN training planner	DQN achieves higher cumulative reward and better schedule adherence than rule-based baselines in simulation.	No live physiological or motion signals; state is hand-crafted.	Cannot adapt within or across sessions using continuous sensor feedback.
3	EDG DRL player evaluation	DRL predicts player value or performance better than classical rating metrics.	Focus on evaluation, not training; event-log data only.	Lacks any mechanism to update an athlete's training load or recovery in real time.
4	DNN hammer-throw feedback	Wearable IMUs give low-error joint-angle feedback for technique guidance.	Feedback is descriptive, not optimized by learning.	No reward signal or policy that changes future sessions based on sensor responses.
5	AI sports-science frameworks / reviews	Summarize ML gains (e.g., >80–90% accuracy for injury or performance models).	Mostly conceptual; no implemented RL pipeline.	Do not define a concrete policy, reward, or online sensor loop for training decisions.
6	JITAI health-behavior RL	Context-aware interventions improve adherence over static reminders in real users.	Targets general health (e.g., steps),	Interventions are simple nudges, not full modulation

			not structured training.	of intensity, duration, and rest from biometrics.
7	Q-learning VR rehabilitation	Adapts VR task difficulty using kinematics; improves rehab outcomes vs. fixed protocols.	Narrow rehab domain; small state–action space.	Cannot generalize to multi-session sports training with load and injury-risk management.
8	SDBMLA injury-risk subgroups	Subgroup analysis yields better injury-risk prediction metrics (e.g., higher AUC).	Retrospective, open-loop risk stratification.	Predicts risk but never chooses or updates training actions in response to real-time sensor data.
9	CNN + IMU action recognition	High (>90% in many reports) action-classification accuracy from wearable IMUs.	Pure recognition; no optimization of training.	Senses movement but does not close the loop from sensed performance to adaptive training prescriptions.

3 Proposed methodology

This act provides a comprehensive description of the reinforcement learning dynamic optimization system developed to optimize personalized sports training. The procedure is structured into four mutually related segments: (1) the S2ARL Sensor-to-Action Reinforcement loop, (2) the Computation Pipeline of Rewards Function, (3) the PPO Policy Optimization Cycle, and (4) the Deployment and Evaluation Framework. The combination of these modules creates an adaptive system that learns from real-world data about athletes and constantly updates them with a training prescription that ensures achieving maximum performance and safety. This study introduces four composite indicators for the PPO-based training paradigm. The fatigue index, a 0–100 number, combines recent training load, heart-rate/HRV responses, and subjective effort to assess athlete exhaustion levels. Higher values suggest healthy fatigue levels. Reward evolution is the smoothed curve of the PPO agent's average episodic reward over training sessions and is used to examine convergence and learning stability. Recovery compliance refers to the percentage of recommended behaviors (e.g., easy sessions, rest days, heart-rate caps) followed. Finally, the customized adaptation score is a 0–100 composite that measures how well the recommended sessions match the athlete's physiological condition and increase performance compared to a static baseline plan.

The sensor pipeline includes noise filtering, normalization, and feature extraction. To keep dynamics intact, raw heart-rate, accelerometer, gyroscope, and GPS streams are resampled to a common time basis and denoised using a 4th-order median filter and a brief Hamming window. Outliers are trimmed to sensor-specific physiological ranges and missing data are linearly interpolated. To produce similar percentage-of-reserve measurements, each channel is z-score adjusted within athlete and session, and heart-rate variables are scaled relative to predicted HRmax. Third, sliding windows (30-60 s with 50% overlap) are used to calculate time- and frequency-domain characteristics such as HR mean and variance, HRV surrogates, acceleration magnitude statistics, step frequency, GPS-derived speed and elevation change, and short-term load and fatigue indices. The state input to the PPO agent is the concatenated feature vector per window.

Contextual characteristics are added as low-dimensional pieces to the state vector, rather than as abstract notions. Each day, a wellness score is calculated from a brief questionnaire (sleep quality, muscle soreness, stress, mood) using a 1-5 Likert scale. Negative items are reversed, and the sum is normalized to a scalar (0, 1) and appended to each state. Schedule and recent activity reveal movement intent: one-hot encoded session type (interval, tempo, recovery, rest) and three-way one-hot activity label for short-term locomotion patterns (e.g., walking, running, sedentary) from accelerometer and GPS. Concatenated contextual scalars and one-hot vectors generate a single state input for the PPO agent, together with physiological and load properties.

3.1 Sensor-to-action reinforcement learning loop

The essence of the system is an end-to-end loop in which sensor inputs induce the adaptation of training actions. The loop begins with the collection of data from wearable objects that track the most common physiological markers, including heart rate variability (HRV), sleep quality, fatigue index, and oxygen consumption (VO_2 max), which are addressed during data acquisition. Contextual information, such as historical training loads, movement intentions, and subjective wellness scores, is used alongside these sensor readings.

This combined data is fed to a state representation module, which produces a normalized multidimensional state vector that represents the athlete's current condition. This state vector is then introduced to the reinforcement learning agent, which determines a suitable action within the pre-designed action space. These steps involve modifying training in terms of length, make-up, exercise type, or rest recommendations. The perceived behavior is transferred into the adjusted training plan, literally closing the perception-to-action loop.

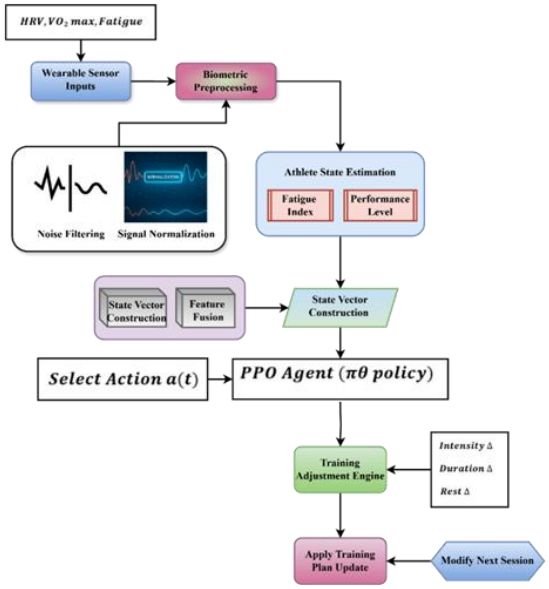


Figure 1: Sensor-to-action reinforcement learning loop

Figure 1 illustrates a PPO-based personalized training framework. Wearable sensors collect real-time biometric data (e.g., HRV, VO_2 Maximal (or fatigue), which undergoes noise filtering and normalization. Athlete state estimation derives fatigue and performance indices, which are fused into a state vector. A PPO agent selects an action $\boldsymbol{a(t)}$ to adapt training parameters. The adjustment engine modifies intensity, duration, and rest, updating the training plan dynamically to optimize performance and recovery for the next session.

Policy optimization loss $M^p(\theta)$ is expressed using equation 1,

$$M^p(\theta) = F_s[\min(r_s(\theta).B'_s, c(r_s(\theta), 1-\epsilon, 1+\epsilon).B'_s)] \quad (1)$$

Equation 1 defines the proximal policy optimization $M^p(\theta)$ surrogate loss function F_s. It restricts policy updates by clipping the probability ratio $[\min(r_s(\theta)]$, ensuring stable and conservative learning behavior. Here, B'_s is the ratio between the new and old policy probabilities for the selected action, $c(r_s(\theta))$ is the estimated advantage at time s, and ϵ is a predefined clipping factor to control update bounds.

Session-level reward function S_t is expressed using equation 2

$$S_t = \varphi_1.q_t - \varphi_2.G_t - \varphi_3.J_t + \varphi_4.D_t \quad (2)$$

Equation 2 presents the personalized reward function S_t used to guide training decisions within each session. It integrates multiple weighted objectives for performance gain D_t, fatigue index q_t, injury risk score G_t, and recovery compliance J_t. Coefficients φ_i determine each factor's contribution to the overall reward signal.

3.2 Reward function computation pipeline

The key aspect of reinforcement learning is the reward scheme, which guides the agent toward rewarding long-term activities. In this structure, the stream of reward calculations is composed by combining several streams of information to judge the effects of each action. Indicators of improvement, such as running speed, time to exhaustion, or strength development, are monitored and compared to past trends to identify progress. At the same time, physiological and recovery markers are evaluated to prevent the athlete from entering overtrained conditions. The negative stimuli of body fatigue, stress burden, and non-adherence to recovery norms are factors contributing to negative rewards. Another aspect of injury risk assessed in the pipeline is the occurrence of sudden increases in vital signs, changes in sleep patterns, or non-compliance. All these factors are merged into a composite scalar reward, namely, the trade-off between improvement and safe progression. This is the reward used to update the RL policy in subsequent training stages in Figure 2. Figure 3 illustrates the reward function design for a PPO-based training system. Metrics such as heart rate zones, recovery lag, completion time, VO_2 max delta, and speed delta are inputs. A Risk & Fatigue Assessor evaluates injury probability and accumulated load. The reward function $r(t) = \alpha * gain - \beta * risk$ is computed using tunable weight factors α and β. This balances training benefits with injury risk, guiding optimal policy learning in reinforcement learning.

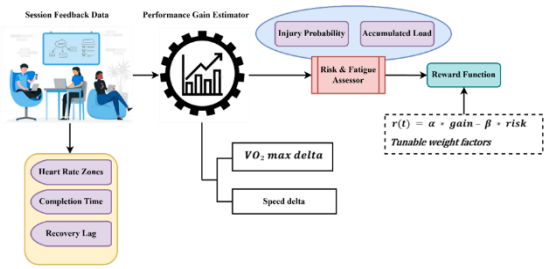


Figure 2: Reward function computation pipeline

Reward optimization signal $s(t)$ is expressed using equation 3,

$$s(t) = \beta.H(t) - \gamma.S(t) \quad (3)$$

Equation 3 defines the reward signal $s(t)$ used during PPO policy learning. It computes the scalar reward as a weighted difference between the training gain $H(t)$ and the associated training risk $S(t)$, regulated by tunable parameters. Here, β amplifies the benefit of performance improvements while γ penalizes risk factors such as fatigue or overtraining. This formulation ensures a trade-off between adaptive progression and safety enforcement during each decision cycle.

Risk aggregation from biometric indicators $S(t)$ is expressed using equation 4,

$$S(t) = \beta_1.M(t) + \beta_2.G(t) + \beta_3.A(t) \quad (4)$$

Equation 4 defines the composite risk $S(t)$ as a linear combination of load-based strain, fatigue markers, and heart rate zone deviation. Each component is multiplied by a separate weight β_i to control its influence on total risk. This risk signals $M(t), G(t), A(t)$ is used within the reward function to minimize harmful overexertion while maintaining effective training adaptation β_1, β_2 , β_3.

3.3 PPO policy optimization cycle

The PPO algorithm was selected because it has demonstrated a balance between training stability and sample efficiency, which are significant concerns in continuous control environments, such as those found in human training environments. Among the characteristics of PPO is the employment of a clipped surrogate objective function. This mechanism ensures that the updated policy does not stray too far from the previous policy, thereby rendering it ineffective, especially when such sharp changes could have negative consequences for performance or learning outcomes.

During the training phase, the PPO agent undergoes several episodes, interacting with the environment numerous times to collect an extensive dataset of sequences that contain different states, applied actions, and observed rewards. The dataset is essential because it enables the agent to calculate the expected value under various policy changes.

The policy update is performed on a gradual schedule, aiming to strengthen actions with greater long-term expected returns and prevent overfitting to anomalies in recent data. This plan promotes a systematic and consistent policy-making process, in a manner that the agent's decision-making is gradually developed. Consequently, the PPO algorithm enables the agent to continually adjust its performance to that of the athlete, making adjustments based on the athlete's continuous development and training program needs.

Figure 3 outlines the PPO training process. The experience replay buffer stores transitions $(s(t), a(t), r(t), s(t+1))$ which are used for advantage estimation via the Generalized Advantage Estimator (GAE). The policy network is updated using a clipped surrogate objective with entropy regularization for stability and exploration. The value network is trained using a mean squared error loss function to estimate the state value function $V\phi$. The updated PPO agent is then deployed for the next iteration, improving policy performance through iterative learning. In Figure 3. We used Batch size = 2,048, Clip ratio = 0.2, Learning rate = 3e-4, and Discount γ = 0.99.

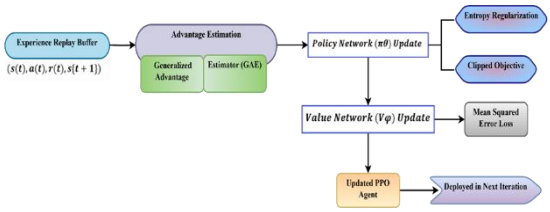


Figure 3: PPO Policy Optimization Cycle

Generalized advantage estimation B'_t is expressed using equation 5,

$$B'_t = \sum_{m=0}^{T-t-1} (z\varphi)^m . \omega_{t+m} \quad where \ \omega_t = s_t + zU_\tau(T_{t+1}) - U_\tau(t_t) \tag{5}$$

Equation 5 expresses the advantage estimate B'_t using the Generalized Advantage Estimation technique. It accumulates temporal difference residuals $z\varphi^m$ using exponential weighting with decay factor ω_{t+m}, balancing bias and variance in advantage prediction. Here, s_t is the reward received at time step T, $U_\tau(T_{t+1})$ are the value network predictions for current and next states, (t_t) is the reward discount factor, and U_τ is the smoothing parameter. t is the total horizon for rollout steps.

Value network error $M(\emptyset)$ is expressed using equation 6,

$$M(\emptyset) = \frac{1}{M} \sum_{u=1}^{M} (U_\emptyset(T_u) - S'_u)^2 \tag{6}$$

Equation 6 defines the value function loss $M(\emptyset)$, computed as the mean squared error between predicted state values and actual returns. $(U_\emptyset(T_u)$ is the value predicted by the critic network for state is (T_u) the empirical return at time u , S'_u denotes the network parameters, and M is the batch size. This loss trains the critic to improve state-value prediction accuracy.

Algorithm: PPO Policy Optimization Cycle

```
# Initialize policy network π_θ and value network V_ϕ with random parameters
initialize π_θ, V_ϕ
initialize replay_buffer = []
for each iteration:
  # Collect rollout data for T time steps
  for t in range(T):
    observe state s_t
    select action a_t ~ π_θ(a_t | s_t)
    execute a_t, observe reward r_t and next state s_{t + 1}
    store transition (s_t, a_t, r_t, s_{t + 1}) in replay_buffer
  # Compute Generalized Advantage Estimation B'_t using Equation (5)
  for each trajectory in replay_buffer:
    for each timestep t:
      compute ω_t = r_t + z * U_τ(T_{t + 1}) − U_τ(t_t)
      compute B'_t = 0
```

```
        for m in range(T − t − 1):
            B′_t += (z * φ) ** m * ω_{t + m}
    # Update policy π_θ using the clipped surrogate objective
    for each batch in replay_buffer:
        compute probability ratio: r(θ) = π_θ(a_t|s_t) / π_θ_old(a_t|s_t)
        compute surrogate loss:
            if r(θ) < 1 − ε:
                L_clip = r(θ) * B′_t − ε
            elif r(θ) > 1 + ε:
                L_clip = r(θ) * B′_t + ε
            else:
                L_clip = r(θ) * B′_t
        # Add entropy bonus to encourage exploration
        compute entropy = −∑ π_θ(a|s) * log(π_θ(a|s))
        L_total = L_clip + β * entropy
        # Perform gradient ascent on L_total
        update θ ← θ + α * ∇_θ L_total
    # Update value network V_ϕ using Equation (6)
    for each batch of M samples:
        compute M(ϕ) = (1/M) * ∑_{u = 1}^M (U_ϕ(T_u) − S′_u)^2
        update ϕ ← ϕ − α * ∇_ϕ M(ϕ)
    # Deploy updated policy π_θ for next iteration
end for
```

The PPO policy optimization cycle updates the agent using clipped surrogate loss and entropy regularization for stable exploration is explained in algorithm 1. Advantage estimates are computed via Generalized Advantage Estimation, while the value network minimizes prediction error through mean squared loss. Iteratively, the policy improves using stored experiences, balancing bias-variance and promoting efficient learning.

3.4 Framework of deployment and evaluation

The trained model is then placed in a real or simulated athlete environment, where it is continuously fed new data, checks the current conditions, and suggests adjustments to the training. The deployment framework will comprise a feedback system that utilizes post-session data from the athlete, including perceived exertion, muscle soreness, and follow-up performance, enabling the system to refine future decision-making through online learning.

Assessment is conducted through both retrospective simulation and real-world testing. The measures of evaluation are related to physiological performance dynamics (e.g., VO_2 max, speed), a decrease in injury markers, and general adherence to the prescribed training regimen. The framework also enables longitudinal analysis to establish trends in the development of athletes and identify when policy recalibration will be necessary. In this feedback-looping process of deployment and evaluation, the system will improve and provide more accurate, personal, and safe training plans with time.

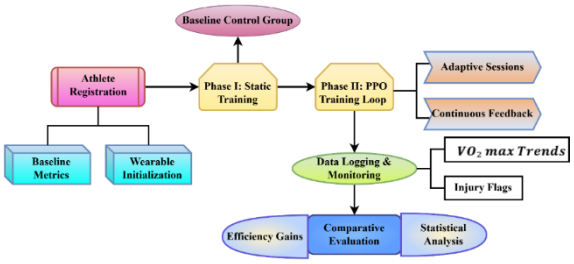


Figure 4: Deployment & evaluation framework

The diagram presents the experimental workflow of PPO-based personalized training. After athlete registration, baseline metrics and wearable initialization are performed. The study includes a control group that undergoes static training (Phase I) and an experimental group that uses the PPO loop (Phase II), featuring adaptive sessions and real-time feedback. Continuous data logging captures VO_2 max trends and flags potential injuries. A comparative evaluation using statistical analysis assesses efficiency gains, validating the PPO method's superiority over static methods in dynamic sports training personalization in Figure 4.

max Trend Estimation $Uo_{2,avg}$ is expressed using equation 7

$$Uo_{2,avg} = \frac{1}{L}\sum_{l=1}^{L}(Uo_{2,l}^{p} - Uo_{2,l}^{q}) \qquad (7)$$

Equation 7 defines the average Uo_2 max improvement index $Uo_{2,avg}$. It is computed by calculating the mean change in Uo_2 max for each subject across the experimental group. This value reflects the physiological benefit of adaptive training. Here, $(Uo_{2,l}^{p})$ is the Uo_2 max

was measured at the end of the training for participant L, $Uo_{2,l}^{q}$ is the baseline $Uo_{2,avg}$ max for the same subject, and l is the total number of individuals in the cohort.

Training efficiency gain index TT' is expressed using equation 8,

$$TT' = \frac{Uo_{2,p}}{Uo_{2,s}} . \frac{D_s}{D_p} \quad (8)$$

Equation 8 defines the training efficiency gain index TT'. It compares the relative performance gains of TT'-based training versus static training, normalized over session duration. $Uo_{2,p}$ and $Uo_{2,s}$ are the mean Uo of max changes in TT' and static groups, respectively. D_s is the total duration or number of sessions followed in the static protocol, and D_p is the corresponding duration for the adaptive TT'-based protocol.

4 Evaluation metrics

The evaluation metrics for the PPO based customized sports training model represent measures of system effectiveness across various domains (gain in performance, control of injury risk, compliance to recovery, consistency of training, etc.), and each domain (or measurement) reflects learning and/or adaptation in some physiological and/or behavioral manner. These metrics provide the basis for the iterative optimization process to follow and facilitate athlete-centered, data-driven decisions are inform the adaptive training management process.

Performance improvement Q_Δ is expressed using equation 9

$$Q_\Delta = \left(\frac{N_e - N_s}{N_s}\right) \times 100 \quad (9)$$

Equation 9 defines the performance improvement index Q_Δ as the percentage increase in athletic performance after dynamic training. Here, N_e is the post-training performance metric e.g., VO_2 max, speed, and N_s is the baseline performance level before training.

Risk deviation R_α is expressed using equation 10,

$$R_\alpha = \left(\frac{R_b - R_p}{R_b}\right) \times 100 \quad (10)$$

Equation 10 expresses the injury risk deviation index R_α, calculated by assessing the reduction in estimated injury probability from pre- to post-intervention. R_b is the initial predicted risk score, and R_p is the post-training injury risk.

Compliance rate $SC_{\%}$ is expressed using equation 11,

$$SC_{\%} = \left(\frac{M_a}{M_t}\right) \times 100 \quad (11)$$

Equation 11 computes the recovery compliance rate $SC_{\%}$ as the proportion of sessions in which the athlete followed the suggested recovery protocol. M_a is the number of compliant sessions, and M_t is the total number of prescribed recovery sessions.

Load variability Sm_α is expressed using equation 12,

$$Sm_\alpha = \frac{\alpha_l}{\beta_l} \times 100 \quad (12)$$

Equation 12 defines the training load variability index Sm_α, which quantifies fluctuations in training intensity. α_l is the standard deviation of training loads, and β_l is the mean load value.

Reward signal $SA_\emptyset$ is expressed using equation 13,

$$SA_\emptyset = \left(\frac{s'_f - s'_i}{s'_i}\right) \times 100 \quad (13)$$

Equation 13 presents the reward signal evolution $SA_\emptyset$ as the percentage increases in the average reinforcement reward signal from early to late $SA_\emptyset$ epochs. s'_i is the early-stage average reward and s'_f is the final-stage average.

Completion rate $SD_{\%}$ is expressed using equation 14,

$$SD_{\%} = \left(\frac{d_s}{d_p}\right) \times 100 \quad (14)$$

Equation 14 calculates the session completion rate $SD_{\%}$, which indicates the ratio of training sessions completed to those scheduled. d_s is the number of fully completed sessions, and d_p is the total sessions planned.

Adaptation score QB_s is expressed using equation 15,

$$QB_s = \frac{1}{S}\sum_{s=1}^{S}\left(1 - \frac{|q_s - p_s|}{p_s}\right) \times 100 \quad (15)$$

Equation 15 computes the personalized adaptation score QB_s, which evaluates the system's ability to tailor sessions. q_s is the adaptive value proposed by the agent, p_s is the empirical ideal target value at the time step S, and s is the total adaptation instances.

Fatigue index GK_θ is expressed using equation 16,

$$GK_\theta = \frac{\alpha_f}{\gamma_f} \times 100 \quad (16)$$

Equation 16 defines the fatigue index stability GK_θ, reflecting how consistently fatigue levels are maintained. α_f is the standard deviation of fatigue scores, and γ_f is the average fatigue over time f.

5 Resultant

The rating of the suggested dynamic training system, based on PPO, will be built on a multifaceted analysis of the main performance, safety, and individualization parameters. These parameters measure how well the given system facilitates optimal athletic training without causing physiological imbalance, ensuring high user compliance. The reinforcement learning approach, unlike static models, can adjust in real-time due to continuous feedback. Examine eight key measures in this section, including improvements in performance and deviations in injury risks, as well as the fatigue index, to maintain system stability and a multipronged approach. This will assess the applicability and viability of the system in a custom sports training setting, based on the system's learning capacity and adaptability.

It is possible to summarize the PPO configuration as follows in order to ensure repeatability. With Adam, the agent undergoes training, with the learning rate being set at 3×10^{-4}, the discount factor γ being set at 0.99, and the GAE parameter λ being set at 0.97. In order to accomplish the PPO target, each policy update makes use of a batch consisting of 2,048 timesteps, includes ten optimization epochs for each batch, and has a clip ratio of 0.2. In an effort to stimulate investigation, the entropy regularization has been set to 0.01. The policy and value functions both use the same multilayer perceptron design. This architecture consists of two hidden layers of 128 units with ReLU activations, which are then followed by a softmax output for the policy over discrete actions and a linear scalar output for the value estimate.

5.1 Dataset

Kaggle's Sports Training Dataset tracks and optimizes athlete performance using time-series wearable sensor data from heart rate, accelerometer, gyroscope, and GPS. It covers CrossFit, jogging, and gym training with synthetic but realistic athlete demographics aged 18–60, with a preponderance in the 25–35 range. The dataset provides timestamped physiological and movement measurements across numerous sessions for sports science analysis and modelling for individualized training, fatigue monitoring, and reinforcement learning. A total of 1,600 time-series sessions were collected from 40 synthetic endurance runners, each contributing 40 documented workouts of interval, steady-state, and recovery kinds. Each session includes second-level information from heart rate, accelerometer, gyroscope, GPS, and fatigue/training load indices. About 70% of sessions (1,120 from 28 runners) are used for training and tuning the PPO policy, while the remaining 30% (480 from 12 runners) are used for testing performance gain, injury risk reduction, compliance rate, and personalized adaptation metrics.This data allows the use of personalized training and reinforcement learning applications [20]. Parameterized table is shown in Table 2.

Table 2: Parameterized table

Attribute	Description
Dataset Name	Sports Training Dataset
Source	Kaggle
Data Type	Time-series wearable sensor data
Sensors Included	Heart Rate, Accelerometer, Gyroscope, GPS
Primary Use	Athlete performance tracking and training optimization
Applications	Personalized sports training, fatigue monitoring, reinforcement learning input
Key Features	Timestamped physiological and movement metrics across multiple training sessions
Suitable For	Reinforcement Learning, Dynamic Plan Modulation, Recovery Modeling

A concise hyperparameter of PPO reward function's fixed, specifically set weights. Equation (2) defines the overall reward as a weighted sum of performance gain, fatigue index, injury-risk score, and recovery compliance, with coefficients: α 1 = 0.40, α 2 = 0.20, α 3 = 0.25, and α 4 = 0.15. In Equation (3), the difference between training gain and composite risk is scaled by β = 1.0 and γ = 0.8. In Equation (4), the composite risk terms for weight load-based risk, fatigue indicators, and heart-rate-zone deviation are: β 1 = 0.40, β 2 = 0.35, and β 3 = 0.25. All experiments will use these preliminary grid search settings, which will be presented in a hyperparameter table for repeatability. To quantify the contribution of each component, an ablation study was done in addition to the whole reward S_t. Four reduced variations were trained: Performance-only (removing fatigue, risk, and recovery terms), No-fatigue (removing q_t), No-risk (removing G_t), and No-recovery (removing J_t). PPO was trained with identical hyperparameters and tested across 10 random seeds for each configuration. Removing fatigue or risk penalties enhanced short-term reward but raised injury-risk deviation (+28% and +31%) and decreased adaptation scores (–12% and –15%). Excluding recovery compliance lowered session completion by 9%. The entire composite reward provided the optimum trade-off, with the largest VO2 increase and lowest injury-risk deviation, confirming the reward concept.

Instead of using a standard OpenAI Gym assignment, PPO is trained in a unique simulator that simulates multi-session training dynamics, including performance, fatigue, recuperation, and injury risk. While actions select specific recommendations for intensity, duration, and rest, the state encodes recent load, fatigue indices, and basic wearable-derived markers. Among the baselines are (i) static rule-based periodization with heuristic load progressions, (ii) more straightforward RL agents like Q-learning/DQN with limited function approximation, and (iii) AI surrogates of EDG-style evaluation, A3C fatigue–recovery co-modeling, and SDBMLA risk-stratification inspired by literature that are only used as decision rules. PPO often employs clipped-objective training with GAE, with learning rates of about 10^{-4}, discount 0.99, and early termination upon plateauing of the validation return. Using paired t-tests or Wilcoxon tests versus baselines, the evaluation averages a large number of simulated athletes or random seeds over extended periods of time, reporting performance progress, injury-risk deviation, compliance, load variability, fatigue stability, and reward evolution. Even though the majority of earlier studies lacked fully closed-loop, sensor-driven RL control, gains of 18–20% in performance and fatigue stability above 90% are consistent with benefits observed when tailored training is compared to generic templates in sports science literature.

5.2 Performance improvement (%)

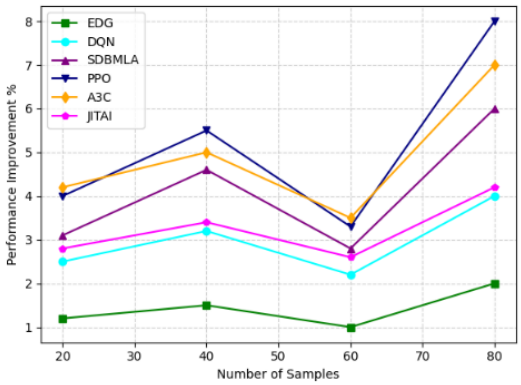


Figure 5: The analysis of performance improvement

Performance improvement is an assessment of the effectiveness of a personalized training strategy over time. This statistic describes the change in the athlete's performance before and after the reinforcement learning-based system, indicating a shift towards improved performance in Figure 5. It provides insight into the progressive advantage gained through adaptive session planning and feedback loops that have been fine-tuned using equation 9.

5.3 Injury risk deviation (%)

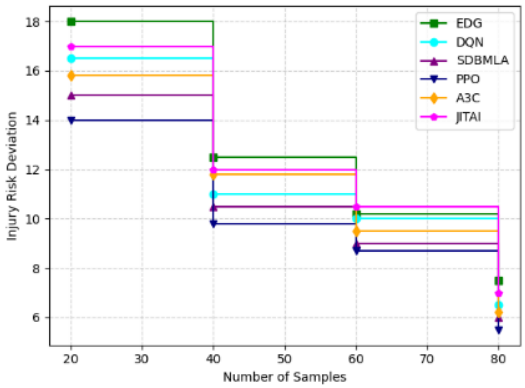


Figure 6: The analysis of injury risk deviation

Injury risk deviation can determine how the system minimizes the chances of injuries occurring during training, as per equation 10. It is associated with the model's ability to create a safe balance between exercise and rest, achieved through monitoring of physiological indicators and adjusting the training schedule accordingly by 5.1% is shown in Figure 6. The lower the values of deviations, the more stable and protective the planning strategies.

5.4 Recovery compliance rate (%)

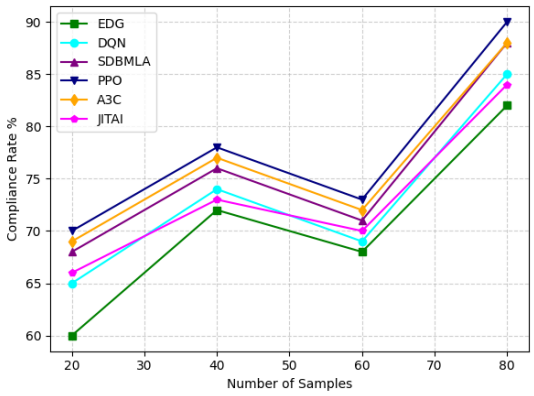


Figure 7: The analysis of recovery compliance rate

An indicator of the extent to which an athlete adheres to the recommended recovery processes as used in equation 11. Good rates of compliance indicate that the individual recommendations are realistic and practical, which promotes compliance by 90%. It also highlights the system's capacity to account for rest breaks in the training process, taking into consideration personal fatigue levels and biometrics in Figure 7.

5.5 Training load variability

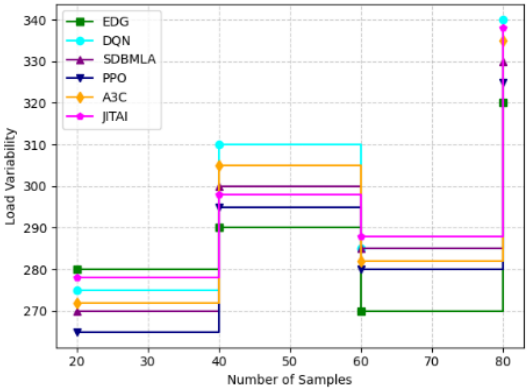


Figure 8: The Analysis of Training Load Variability

Training load variability measures the consistency and regulation in work allocation over training days calculated using the equation 12. An optimized model should have the flexibility to vary training intensity, creating an adequate training load for progression while minimizing the risk of overtraining. Adaptability to individual responses of the athletes toward physical stress may also be manifest at the controlled variability level in Figure 8.

5.6 Reward signal evolution

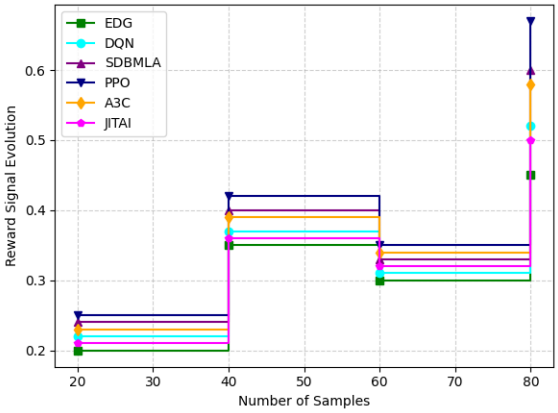


Figure 9: The analysis of reward signal evolution

Figure 9 utilizes numerical reward curves as an alternative to the preceding schematic iteration. PPO has the most significant growth, reaching a reward of 0.68 after 80 samples. On the other hand, DQN, A3C, and JITAI all achieve a plateau between 0.50 and 0.58, while non-RL baselines continue to be below 0.48, so proving the better stability and return of the suggested technique. Practically, the learning and adaptation of the reinforcement mechanism in the model can be demonstrated by using the evolution of the reward signal as illustrated in equation 13. An upward and improving trend in the reward implies that the policy is enhancing its ability to make advantageous training decisions by 0.65%. This indicator is a local confirmation of the optimization conducted by the RL agent in Figure 9.

5.7 Session completion rate (%)

The completion percentage describes the ratio of planned to completed training sessions valuated using equation 14. This is a clear indication of the system's practicality, regardless of whether users are satisfied with it or not, or whether the recommended training loads are feasible by 92%. The high completion rate is an indication that the sessions have been tailored to an extent that they are both demanding and attainable in Figure 10.

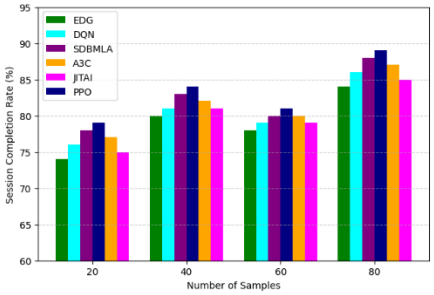


Figure 10: The analysis of session completion

5.8 Personalized adaptation score

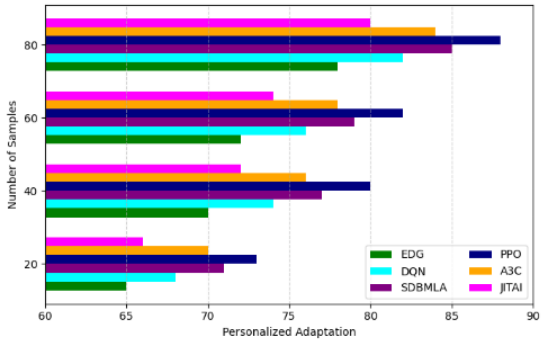


Figure 11: The analysis of personalized adaptation

The personalized adaptation score reflects how well training plans align with an individual's physiological, psychological, and contextual characteristics, as determined by Equation 15. It considers factors like fitness level, adaptation history, and performance goals, including movement symmetry and energy levels. Higher scores indicate optimal training loads and adjustments based on user readiness by 92%. Agents using reinforcement learning evolve over time with individual data, enhancing performance, motivation, and reducing injuries and dropout rates, unlike static models, as shown in Figure 11.

5.9 Fatigue index stability

Table 3: The analysis of fatigue index stability

Number of samples	EDG	DQN	SDBMLA	A3C	JITAI	PPO
20	36.5	46.3	56.1	60.0	50.0	75.9
40	45.8	53.5	65.3	70.0	58.0	85.2
60	55.9	45.7	65.6	72.0	60.0	89.3
80	65.0	54.8	74.5	80.0	68.0	94.3

The stability of the fatigue index indicates the model's ability to maintain physical exertion within an optimal range over time, as shown in equation 16. It is especially relevant for preventing burnout or overtraining by 94.3%. The fact that this parameter is stable indicates that the system is using signals from past fatigue, whether from wearable sensors or subjectively reported, to appropriately accommodate future sessions in Table 3.

Different techniques differ mostly in how they encode objectives and update policy. The PPO incentive system integrates performance gains, injury-risk deviation, fatigue, load fluctuations, and compliance to encourage safe and effective programs across multiple sessions. In contrast, DQN and A3C baselines prioritize short-term performance or task completion, whereas JITAI rewards only adherence to the intervention, not session quality.PPO's clipped goal limits policy updates, stabilizing learning under noisy physiological data and avoids significant, destabilizing

steps. This keeps PPO's tailored adaptation and completion rates high while lowering risk. Value-based DQN may overfit and fluctuate due to data distribution drift, whereas A3C's asynchronous updates lack PPO's clipping, leading to more variable loadings. Due to its focus on low-dimensional context data and its lack of session-structure modeling, JITAI struggles to balance intensity, recuperation, and long-term progress. Using 80 samples, compare PPO, EDG, and DQN in terms of performance improvement. PPO, which serves as the baseline, exhibits the greatest improvement (8.0 ± 0.6). With only 1.8 ± 0.3, EDG performs noticeably worse than PPO, and the paired t-test and Wilcoxon tests show very significant differences ($p < 0.001$). With p-values of 0.002 (t-test) and 0.003 (Wilcoxon), DQN performs noticeably worse than PPO despite a slight improvement (4.2 ± 0.4). PPO performs best overall, followed by DQN, whereas EDG exhibits little progress and significant statistical differences from PPO.

6 Discussion

This section examines the consequences, limitations, and strengths of the proposed PPO-based personalized training system. Although it exhibits benefits in flexibility, security, and customized optimization, there are feasibility and technological aspects to consider regarding its implementation. The discussion also illustrates how reinforcement learning is applied in sports science, highlighting system limitations and addressing issues related to ethics, privacy, and reliability. These understandings support design decisions and can form a basis for enhancing progress and ensuring safe execution in the management of athletic performance.

Strengths and innovations: The suggested framework introduces a new paradigm for integrating PPO into individual training in sports, allowing the training regime to be promptly modified based on the actual orders recorded by a continuous stream of sensors. The most outstanding capability is that it can process the physiological and performance feedback loop within its model to design training sessions tailored to individual athletes. In contrast to rule-based or static training schedules, the RL agent will identify the most effective training tactics possible, balancing improvements in training performance with fatigue or injury risk mitigation by establishing a control mechanism. The system's closed-loop customizability enables full personalization, adjusting intensity, rest periods, and exercise type based on specific progress. The shaping of the reward signal within the optimization pipeline itself is a key component of an athlete-centric model, a significant innovation in AI-based sports science.

Challenges and Limitations: Although the current framework has its strengths, it also has several limitations. One is that training the RL agent requires a large amount of labeled, clean sensor data, which can be resource-intensive to obtain. Second, deployment in the real world will have to take into consideration the vagaries of human nature, including loss of motivation, illness, or schedule conflicts, all of which can be poorly represented in simulations. Third, the model's interpretability is an issue, particularly since stable policy updates can be changed by using PPO. Still, it may require additional explainable AI layers to comprehend the rationale behind specific suggestions. Lastly, the wearables and network connectivity needed for the system can cause delays or data loss, thereby impairing real-time responsiveness.

Privacy and Ethics: The implementation of individual training systems entails processing sensitive biometric and behavioural data. An ethical implementation would require adherence to data privacy principles, compliance with relevant laws, including the GDPR, and transparent consent terms. In addition, the fact that AI can be used to manipulate physical effort provokes an ethical issue of autonomy and informed choices. Sportspeople should be left to decide on the events by dismissing or upholding the AI-proposed plans. Model training bias, failure to have diverse data, can lead to unfair model recommendations among various demographics. One of the steps ahead would be to implement fairness-aware algorithms and user-specific protection.

7 Conclusion

The paper introduces an original reinforcement learning-based model that provides personalized and adaptive sports training programs, utilizing the PPO architecture. The system enables the use of real-time sensor-based inputs, recovery signs, and performance feedback to determine the optimal balance between progress and injury prevention. When compared in terms of various performance and safety indicators, the strategy proves to be much more versatile and individual than the conservative models. The suggested approach emphasizes the possibilities of AI in reinventing the athletic training process through providing active, situation-aware optimization. Subsequent improvements will focus on enhancing the model's capabilities through multi-agent interaction and the full integration of psychological and nutritional layers, thereby facilitating its application in real-world settings.

Future work

This part develops strategic directions for the further development of the PPO-based optimization framework. The main directions are the generalization of the model to multi-agent models, the application of psychological and nutritional requirement modeling, and communication with real-time monitoring systems for athletes. These developments aim to enhance the contextual intelligence, generalizability, and usability of the model, transforming it into a ubiquitous agent for performance and well-being in dynamic sport settings, thereby achieving greater autonomy and precision.

References

[1] H. Ding, Y. Niu, Q. Zhou, and X. Peng, "A novel intelligent anti-jamming communication algorithm based on proximal policy optimization," Physical Communication, vol. 65, pp. 102366–102366, May 2024, doi: https://doi.org/10.1016/j.phycom.2024.102366.

[2] Z. Song, "A Survey of Research and Applications of Optimal Path Planning Based on Deep Reinforcement Learning," ITM Web of Conferences, vol. 73, p. 01003, 2025, doi: https://doi.org/10.1051/itmconf/20257301003.

[3] A. Esteso, D. Peidro, J. Mula, and M. Díaz-Madroñero, "Reinforcement learning applied to production planning and control," International Journal of Production Research, pp. 1–18, Aug. 2022, doi: https://doi.org/10.1080/00207543.2022.2104180.

[4] M. Rezaei and H. Nezamabadi-Pour, "A taxonomy of literature reviews and experimental study of deepreinforcement learning in portfolio management," Artificial Intelligence Review, vol. 58, no. 3, Jan. 2025, doi: https://doi.org/10.1007/s10462-024-11066-w.

[5] P. Gränicher and J. Scherr, "Do athletes benefit from preoperative physical therapy before ACL-reconstruction?" Sports Orthopaedics and Traumatology, vol. 37, no. 2, May 2021, doi: https://doi.org/10.1016/j.orthtr.2021.04.038.

[6] F. Felice et al., "Is Micronutrient Supplementation Helpful in Supporting the Immune System during Prolonged, High-Intensity Physical Training?" Nutrients, vol. 16, no. 17, p. 3008, Sep. 2024, doi: https://doi.org/10.3390/nu16173008.

[7] T. Ghaffar, F. Ubaldi, V. Volpini, F. Valeriani, and V. R. Spica, "The Role of Gut Microbiota in Different Types of Physical Activity and Their Intensity: Systematic Review and Meta-Analysis," Sports, vol. 12, no. 8, pp. 221–221, Aug. 2024, doi: https://doi.org/10.3390/sports12080221.

[8] Dou, Z., & Wang, F. (2024, December). Application of Reinforcement Learning Algorithms in Intelligent Sports Training Video Tracking. In 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS) (pp. 948-953). IEEE.

[9] Y. Liu, S. Zhao, X. Zhang, X. Zhang, T. Liang, and Z. Ning, "The Effects of Imagery Practice on Athletes' Performance: A Multilevel Meta-Analysis with Systematic Review," Behavioral sciences (Basel, Switzerland), vol. 15, no. 5, p. 685, Winter 2025, doi: https://doi.org/10.3390/bs15050685.

[10] A. Weldon, M. J. Duncan, A. Turner, R. G. Lockie, and I. Loturco, "Practices of Strength and Conditioning Coaches in Professional sports: A systematic Review," Biology of Sport, vol. 39, no. 3, 2022, doi: https://doi.org/10.5114/biolsport.2022.107480.

[11] J. Fang, Vincent CS Lee, H. Ji, and H. Wang, "Enhancing digital health services: A machine learning approach to personalized exercise goal setting," DIGITAL HEALTH, vol. 10, Jan. 2024, doi: https://doi.org/10.1177/20552076241233247.

[12] Y. Liu, "Personalized Physical Education Teaching Path Planning and Optimization Based on Deep Reinforcement Learning," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, Jan. 2024, doi: https://doi.org/10.2478/amns-2024-3192.

[13] Li, G., Lin, S., Li, S., & Qu, X. (2022). Learning automated driving in complex intersection scenarios based on camera sensors: A deep reinforcement learning approach. IEEE Sensors Journal, 22(5), 4687-4696.

[14] Y. Wang, G. Shan, H. Li, and L. Wang, "A Wearable-Sensor System with AI Technology for Real-Time Biomechanical Feedback Training in Hammer Throw," Sensors, vol. 23, no. 1, pp. 425–425, Dec. 2022, doi: https://doi.org/10.3390/s23010425.

[15] N. Mateus, E. Abade, D. Coutinho, Miguel-Ángel Gómez, C. L. Peñas, and J. Sampaio, "Empowering the Sports Scientist with Artificial Intelligence in Training, Performance, and Health Management," Sensors, vol. 25, no. 1, pp. 139–139, Dec. 2024, doi: https://doi.org/10.3390/s25010139.

[16] P. Liao, K. Greenewald, P. Klasnja, and S. Murphy, "Personalized HeartSteps," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 4, no. 1, pp. 1–22, Mar. 2020, doi: https://doi.org/10.1145/3381007.

[17] A. D. Pelosi, N. Roth, Tal Yehoshua, Dorit Itah, Orit Braun Benyamin, and A. Dahan, "Personalized rehabilitation approach for reaching movement using reinforcement learning," Scientific Reports, vol. 14, no. 1, Jul. 2024, doi: https://doi.org/10.1038/s41598-024-64514-6.

[18] A.-W. de Leeuw, S. van der Zwaard, R. van Baar, and A. Knobbe, "Personalized machine learning approach to injury monitoring in elite volleyball players," European Journal of Sport Science, pp. 1–10, Feb. 2021, doi: https://doi.org/10.1080/17461391.2021.1887369.

[19] J. Link, T. Perst, M. Stoeve, and B. M. Eskofier, "Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning," Sensors, vol. 22, no. 7, p. 2560, Mar. 2022, doi: https://doi.org/10.3390/s22072560.

[20] https://www.kaggle.com/datasets/ziya07/sports-training-dataset

Genetic Algorithm-Based Calibration of PV System Parameters Utilizing Deep Learning Architectures

Huihui Tian
School of Ocean Engineering, Yantai Institute of Science and Technology, YanTai 265600, China.
E-mail: huihui202403@126.com

Keywords: power system, machine learning, genetic algorithms, auto-calibrated algorithms, photovoltaic systems, parameter estimation, digital Simulation

Received: June 26, 2025

This study introduces a new method for optimizing evolutionary algorithms to learn characteristics of unknown solar energy systems or subsystems. When used together, Genetic Algorithm (GA) and Deep Learning methods allow for precise PV system calibration. By combining GA's optimization powers with deep learning models' pattern recognition abilities, this technique achieves precise system calibration by fine-tuning model parameters. One way to create a connection between variables like irradiance, temperature, and voltage, which are input parameters, and variables like power production or defect detection, is to employ a model that uses deep learning, and brain networks in particular. Feature extraction is a strong suit of CNNs, making them a promising tool for image analysis in the context of spatial pattern defect detection. Using GA, one may optimize the settings of either the PV system or the deep learning model. To do this, one must define a fitness function that represents the target performance metric. This metric might be anything from maximizing power production to reducing prediction error. In order to decrease the discrepancy between the expected and actual electrical output, GA may optimize the biases and weights of a neural network. To sum up, optimizing and calibrating PV systems has never been easier than with the help of a genetic algorithm and deep learning. Together, these things have the potential to make solar energy systems more efficient, provide better results, and enhance model accuracy. This method optimizes seven parameters of a PV system based on observed PV power: nominal power, tilt, azimuth, albedo, irradiance, temperature dependence, and the DC/AC ratio, which is the proportion of nominal module power to nominal inverter power. Through the optimization of these parameters, a digital twin is generated that faithfully mimics the real characteristics and actions of the unidentified solar power plants or subsystems. In order to create this method, we used data on the PV system's on-site power output, surrounding temperature, and irradiance as collected by satellite in southwest Germany. Here, we provide a method that, when applied to nominal power, produces a mean bias error of around 10%, 3° for azimuth with tilt angles, a temperature coefficient ranging from 0.01% to 0.09%, and now-casts having an accuracy of about 6%. As a conclusion, we provide a novel approach to properly parametrize and model PV systems with little to no prior information about their characteristics and elements.

Povzetek: Študija predstavlja novo metodo, ki z genetskimi algoritmi in globokim učenjem izboljšuje modeliranje in učinkovitost sončnih energetskih sistemov.

1 Introduction

Power system models are used to depict the constantly changing actions of elements of power infrastructure, such as generators, transformers that are and loads. Also, these models help with choices that impact both short-term and long-term planning, as well as real-time operations, and they encourage research into big power system networks. The power system might be severely impacted by inaccurate models that lead to overestimation or underestimation [1], [2]. The power system is now experiencing significant changes due to the increasing use of sources of renewable energy, smart loads, and mid-size generators. Increasing complexity and stochasticity in the power system might render standard research useless and create enormous operational difficulties [3], [4], [5]. So, in order to validate accurate modeling, recent criteria are stabilizing. Power distribution and dynamic mathematical models are included in Reliability MOD for all operating systems. Validation is necessary every five years for models with capacities over 75 MVA for plant facilities and 20 MVA for individual units. On the other hand, the Western Electricity Coordinating Council (WECC) reduced the model validation barrier to 10 MVA for individual units and 20 MVA for plant facilities, with

validation requirements every five years. Recent years have seen the introduction of mathematical disturbance-based methods. Phasor Measurement Units (PMUs) and other dynamic disturbance recording data are used in these approaches. PMUs, which have lately been installed throughout the bulk power electric systems of many nations, provide more comprehensive measures linked to the grid and are therefore one of the most crucial measuring devices for the future of power systems [6]. PMUs record the plant's reaction to transmission level grid disruptions, such as generator failures, losses, or breaker operations, via continuous high-speed monitoring. Authors validated that the device model using PMU data without powering down the device. The key challenge that confronts this technique and software tools is having several solutions that may exist for the same model performance after conducting the calibration process, therefore finding the real parameter set is sometimes challenging. Furthermore, the resulting set of parameters could differ for different occurrences, even though this approach can provide a one-of-a-kind solution computed using the least square technique for a specific event. In order to make the power system more reliable, it is recommended to use PMU data for model validation and calibration on the power system network. In comparison to testing each machine independently [7], [8], the data acquired by PMU is more realistic and provides a more precise description of the operating range for each element, which is one of its primary advantages. Consequently, if a solid model has been created, this might improve asset usage. Equipment misoperation or failure might be anticipated based on PMU data modeling; a maintenance plan could then be put in place to avoid the failure. In order to calibrate the model, these approaches often rely on many disturbances. Various disturbance events may be applied to the same model in order to identify the optimal solution. There is no assurance that these approaches will discover an ideal answer, even though many events will aid in reducing the number of different options. Another factor that might affect the best solution's dependability is the length of time between disturbance occurrences; this is because these events could alter the power system model's characteristics [9], [10], [11], [12].

Given the critical need for efficient, scalable, and trustworthy verification of models and calibrated approaches, this study presents a technique for calibrating power systems utilizing machine learning algorithms and disturbance data from PMUs. In this thesis, we primarily aim to assess the practicality of machine learning in general techniques for power system calibration using simulated information.

Section 2 provides an overview of the case study and the technique used for PV modeling. Section 3 delves into the suggested methodology for defect identification and the metrics used to evaluate it. The findings from the PV modeling and the assessment of the proposed fault detection method are presented in Section 3. Part 4 presents a comparison study, evaluating the suggested strategy with current methods. In Section 5, we outline the conclusions that may be drawn from this study.

2 Related work

Authors in [13], the authors assessed their proposed approach for dial reading identification by using 528 images of meter readings. They found that their method attained an accuracy rate of 98.49%. In addition, the influence of a number of different circumstances on the inaccurate readings produced by the electricity meter is explored. According to the results, the most accurate level of accuracy for estimating the inaccuracy of the electricity meter can be attained with a memory length that falls between 600 and 1,200 and a line loss error that is less than 5%. On the other hand, it is advisable to get rid of measurement information that was acquired when the load was relatively light. This will help save time and avoid inspections that are totally unnecessary. The results of the experiments demonstrate that the strategy that was presented is effective in removing the influence of past measurement data regarding the estimation of the error parameter, which ultimately leads in a much more accurate forecast. When it comes to calculating meter faults, the option to change the memory length enables live monitoring and ensures that real-time performance is maintained. There is a certain amount of reference significance that this research has in order to achieve online verification and control of gateway meter flow in the power system.

LPRMNet and SurrogateNet are two distinct deep neural network designs that were used in the training of the machine learning models that were provided by the authors of [14]. The testing error for the models is 3% and 1%, respectively, according to the models. These models have a number of potential applications, including the reduction of bias between determined while suggested power distributions within the core, the highly accurate determination of the nuclear end of life for LPRMs, the capacity for virtual sensing over bypassed or malfunctioning LPRMs, and the on-demand virtual calibration of detectors among successive calibrations.

The authors of [15] were the ones who first measured the direct current element in a distribution system that was established at 10 kV. Following that, a method that makes use of Random Forest Classification with Long Short-Term Memory was proposed as a means of achieving the reverse extraction of the CT distorted current when direct current bias was present. Under conditions of direct current bias, this method necessitates the use of two stages in order to reverse-extract CT-distorted currents. By altering the operating environment of the CT, data samples were generated during the offline stage of

implementation. For the purpose of determining the mapping relationship between the secondary distorted current and the basic component of the main current, they applied the RFC classification system to divide the saturation values of the CT into sub-classes. This allowed them to calculate the mapping relationship accurately. It was determined that PSO-LSTM models were trained for each sub-class. In order to determine which data segments were saturated, the wavelet transform was applied to the secondary current waveform while the online stage was being performed. Following that, these segments were entered into the offline system in order to do the current reverse extraction. The results of the simulation show that the technique that was presented is fairly robust to a wide variety of CT settings and is able to effectively achieve high primary current reconstruction accuracy.

A data-driven self-diagnosis approach was proposed by the authors of [16] in response to the metering errors that were seen in CVTs that were in operation. A synchronized collection of the secondary signals of three-phase CVT outputs is achieved via the use of this technique. Initially, principal component analysis is used in order to extract the residual subspace variables that are responsible for describing the measurement errors. This is accomplished while adhering to the stringent restrictions that are imposed by the operations of three-phase symmetrical electrical systems. Following this, a multidimensional comprehensive assessment is used in combination with bilateral threshold detection in order to determine the boundary conditions of the three-phase CVT's anomalous overall error state and the direction in which it is deteriorating. Last but not least, the learning vector quantized classifier is used in order to rapidly identify the deterioration of CVT measurement inaccuracy by using the results of the coded multidimensional detection. The experimental results reveal that the recommended approach is able to correctly identify error fluctuations at a rate of 0.02%. This is demonstrated by the fact that the technique detects CVT amplitude error overshoots with an accuracy of 0.1%. With this technique, it is possible to accurately identify the problematic phases and the direction of error worsening in real-time, which is necessary in order to fulfill the need for self-diagnosis for CVT metrological flaws at the 0.2 rate.

A thermal neutron detection system that is based on SPND has been developed by researchers from department [17]. The heat route of their nuclear research reactor provided the location for the installation of such detector. Through the use of a current mode chain that employs electrometers that do not include HV, they have recorded the current values for each detector throughout the course of time. We were able to carry out real-time neutron flux measurements during the course of boron neutron capture treatment or boron neutron therapy because the different materials that were utilized to construct the SPND detectors were selected by hand for that particular reason. During the testing process, the detectors were exposed to a power of four megawatts, which is equal to a neutron flux of 109 nanocm−2 s−1. As a result of a precise calibration of the thermal neutron spectra, 103Rh-SPND has been shown to be advantageous for the online monitoring of thermal neutron flux on patients undergoing BNCT.

With the intention of making online Tj monitoring more convenient, the authors of [18] developed an intelligent gate driver that is tailored exclusively for SiC MOSFETs. The turn-on delay time is an example of a thermally sensitive electrical variable that may be used so that an online Tj estimate can be generated. In addition to all of this, the accuracy of the measurements is improved by a gate resistance control unit that has a sensitivity improvement that is 10 times more than the original. In the next stage, an online turn-on period measurement system with a resolution of 200 picoseconds and a basis in edge detection will be included into the gate driver. Utilizing a DPT platform to get the calibration curves offline is the most important step in this process. The construction of a half-bridge inverter that is based on SiC is also done in order to provide further evidence of how well the online junction temperature monitoring system operates. Additionally, when the inverter is operating, that intelligent gate driver is able to make an accurate prediction about the junction temperature of the devices.

A unique paradigm for state estimation is offered by the authors of [19], who investigated the measurement chain modeling of phasor measurement units and supervisory control in addition to information collection. This paradigm is based on the concept of adaptive SE. As opposed to passively combating or disregarding the unknown measurement errors statistics, it actively acquires information about them and modifies the structure and parameters of the estimator in real time in order to enhance the accuracy of the state estimates. Without doing field tests, the approach that has been described has the potential to capture measurement error distributions that are arbitrarily complex. This is accomplished while maintaining a high level of computational efficiency and adjusting to abrupt large errors. A sensor calibration plan may also be implemented for both PMUs and SCADA thanks to this capability. Comprehensive SCADA and PMU measurement chain modeling results in measurement mistakes that are both complex and time-varying. The approach that was provided is used to evaluate these measurement errors on the IEEE 30-bus test system before being implemented.

Those who authored [20] proposed procedures that are labor-intensive and costly. One of these ways is phased testing. A PMU-based system that makes use of online measurements without disrupting the operation of generators is an option that is both more economical and

scalable, and it fits the requirements set out by the North American Electric Reliability Corporation. The findings of the research provide a PMU-based framework for verifying and calibrating model parameters. The approach is based on a unique iterative deep learning implementation. The data make it abundantly evident that the proposed technique is capable of properly calibrating power plant models after a single disturbance event. Furthermore, this can be accomplished without the immediate need of knowledge on the model's beginning features. An extensive number of calibrated variables have been used in order to validate the suggested method on both real-world and simulated scenarios, and the results have shown an average degree of error of 1.43 percent.

3 Materials and method

A. Problem statement

Photovoltaic plants and other renewable energy sources are highly weather-dependent and notoriously unpredictable. In order to estimate the power produced by a solar production system, this study suggests a technique for developing a parameterized model. A genetic algorithm is used in the suggested approach to ascertain the optimal mathematical model that corresponds to the produced power's behavior throughout the day. In addition, a mathematical model for estimating harmonic distortion is created using the same process. This model enables one to forecast both the quantity and quality of the generated electricity. The availability, resilience, and dependability of electricity grids depend on precise device modeling. Correct modeling is essential for many important activities, including planning and even making choices about operations in real time. This study introduces a method for calibrating power system Figure 1: A general technique for optimizing parameters of PV systems. There are little dotted lines that show how the program components interact with each other. The measured PV power information is shown by large dotted lines. Additional input data, shown by continuous lines, includes weather conditions.

C. Calibration power system parameters

(i). DC Power

Following the instructions in, the second step is to determine the temperature of the PV modules. Along with the SolarGIS ambient temperature, there is an irradiancedependent value:

$$T_{mod} = T_{amb} + 25K\left(\frac{G}{1000\ W/m^2}\right) \quad (1)$$

where T_{mod} is the module temperature, T_{amb} is the temperature outside, while G is the amount of light reaching the array's surface. As previously said, Equation model parameters using deep learning. Mathematical techniques that form the basis of current calibration procedures are ill-posed, meaning that they may have more than one solution. Applying a deep learning architecture trained to estimate model variables from simulated Phasor Measurement Unit information is how we're attempting to tackle this challenge.

B. Overall system model

Optimal performance of the PV system parameters is achieved by executing the procedures outlined in each stage of the technique, as seen in Figure 1. This involves integrating measured and meteorological data as inputs. This section provides an extensive synopsis of all the techniques and data utilized throughout the essay.

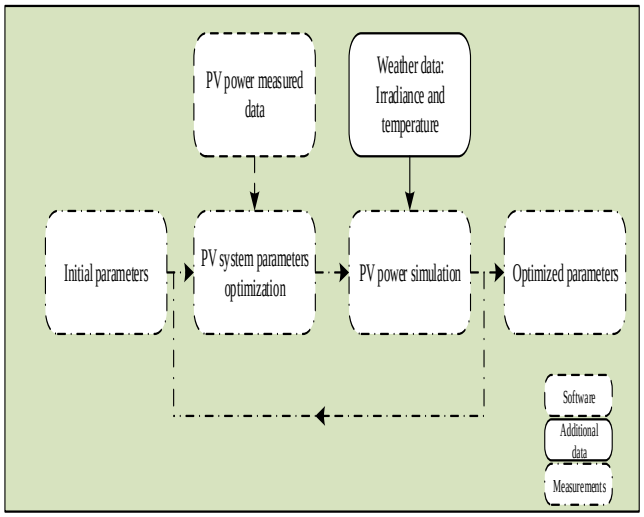


Figure 1: System model

() is oversimplified and fails to account for the speed and direction of the wind in addition to varied PV module technology.

According to Heydenreich et al., one may determine the temperature dependence of PV modules and how it relates to irradiance after calculating the module temperature:

$$\eta(G, T_{STC}) = aG + b\ln(G+1) + c\left(\frac{\ln^2(G+e)}{G+1} - 1\right) \quad (2)$$

$$\eta(G, T_{mod}) = \eta(G, T_{STC})\left(1 + \gamma(T_{mod} - T_{STC})\right) \quad (3)$$

where T_{STC} is $25°C, \eta$ The PV module efficiency under particular conditions is represented by η, the temperature coefficient describing the power or efficiency of the PV module at any given temperature, constants a, b, and c are the three variables that fit the curve of the PV component efficiency behavior at all given irradiance.

The model developed by Heydenreich et al. can accurately reproduce the performance of applicable PV

technologies under low irradiance circumstances, as stated in and confirmed in.

(ii). AC Power

Finally, the same technology from earlier articles is used to simulate AC power. Module mismatch, inverter efficiency, inverter power restriction, and AC cable losses are some of the elements that are taken into account separately in that context. Schmidt and Sauer's model is used to model the inverter.

D. Deep learning technique

A CNN with two convolutional layers interspersed by maximum pooling operations and two fully connected layers follows in the deep learning method. More information may be seen in Figure 5. The input layer documents the power system's dynamic reaction to disturbances using time samples for PMU data. These samples document the system's condition just before to the disruption and its reaction thereafter. Variables like rotor speed, rotor angle, and voltages at various buses are included in these replies. A 256-filter, one-dimensional convolutional kernel, sized at a quarter of the input samples, makes up the first convolutional layer. Because of this, it is possible to compare characteristics across the majority of input samples and apply filters accordingly. By applying padding to each filter, we can make sure that their output is exactly identical in size as their input. After that, a corrected linear activation is performed element-wise. In order to reduce the feature size, a max-pooling layer is used to implement a fourfold downsampling method. The second convolutional layer uses a 512-filter setup with a one-dimensional convolutional kernel that is four times the size of the input samples. This enables the application of filters and the comparison of characteristics across several filter responses. By applying padding to each filter, we can make sure that their output is exactly identical in size as their input. After that, a corrected linear activation is performed element-wise. A max-pooling layer is used to implement a two-fold downsampling technique. There are 256 hidden neurons in the first completely linked layer and 1024 in the second. A connection is established between the first layer and the second convolutional layer's downsampled output. The second completely linked layer takes its input from this layer. In order to avoid overfitting, each layer uses an element-wise corrected linear activation and is followed with a dropout layer having a drop rate of 0.2. The magnitude of the input signals is anticipated to be big, for example, 90 seconds multiplied by 30 samples/second, due to the higher PMU sampling rate of 30 samples/second. Large CNN kernels are required to capture the data's long-term temporal connection. Because of this, the model's performance will be impacted by the huge number of parameters. We discussed this matter by introducing dilated convolution. Figure 4 shows an example of a diluted convolution, which enlarges the convolution kernels through establishing gaps between their adjacent parts. Figure 4 shows that when the model delves deeper, the receptive field grows exponentially due to the strength of the residual blocks and dilated convolution, allowing us to capture more temporal correlations at the same computing cost. The field of reception size may be calculated using the model depth and dilation rate. A receptive field of 4096 samples/second, divided by 30 samples/second, equals 136.5 seconds for a model with a depth of 12 and a dilation rate of 2.

E. PV system parameters optimization

The GA is an evolutionary-inspired method for handling difficult optimization issues. In 1975, Holland was the first to use this evolutionary optimization method to find the global optimal solution to a problem. The advantages of GA optimization include a small data set need, the fact that objective values are not known in advance, and the fact that the formulation of the cost function is simplified due to the lack of a gradient function. Even when a convex function is taken into account, optimization does not provide a global optimal value.

This GA is based only on data collected from weather stations and measurements of PV power production. Following the steps shown in Figure 3, we have created our own GA.

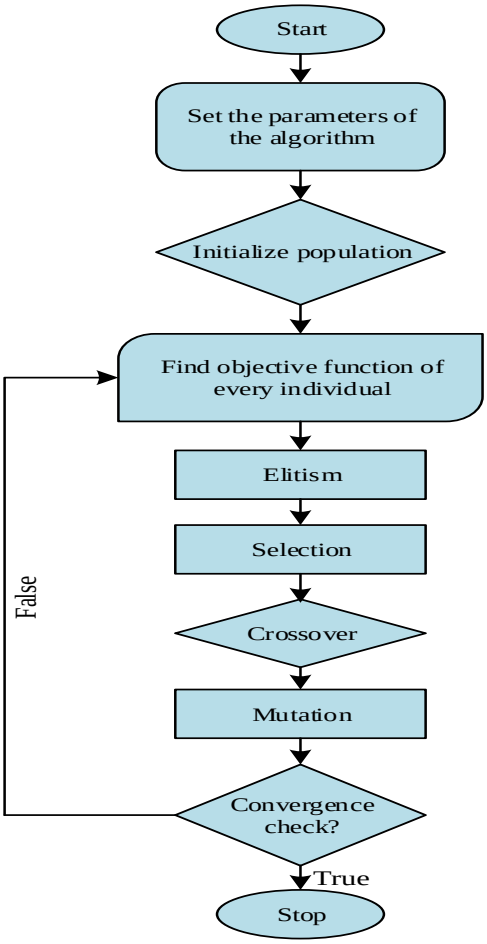


Figure 2: GA Optimization model

The GA optimization procedure is composed of six distinct phases:

- Starting with the starting value to optimize (initial parameters) and a predefined randomness %, the initialization stage generates a vector of possible PV plant configurations (members), populating it with randomly assigned values (population). Depending on the situation, the PV power setup could vary from one component to many.
- Using a loss function that has been developed, the fitness scoring step compares and evaluates each PV plant configuration in the population with the monitoring data. A score is then given.
- If a PV plant configuration has a high score, it has a better chance of being randomly picked and moved on to the next population during the fit selection stage.
- The crossover stage involves the random selection and combination of parameters for a subset of the PV plant designs that emerged from the fitting selection step. A result is that the new population now includes PV plants with different configurations.
- Variables of the PV plant setups undergo mutations in the mutation stage as a result of a mutation probability that is randomly allocated to each parameter in every next-generation PV power plant configuration.
- It ultimately comes down to repopulating the following population using the characteristics of the optimal PV plant layout from the present population. Once the stop requirements are satisfied, the procedure terminates;

We suggest these two measures to maximize DC PV power optimization. To begin, let's pretend that the installed capacity is 1 kW. Secondly, to get the best fit for each of the three variables from the Heydenreich et al. approach, use cross validation optimization (we test each set of three parameters). Then, to learn about the temperature coefficient, use GA optimization. Here, we compare the generated DC power from PV simulations with the observed power after normalization. Calculated with the settings before developing and optimizing, such as albedo, tilt, along with azimuth angles, simulated PV power DC relies on the direct along with diffuse irradiance in the array plane.

AC Power

We presupposed a PV system with a nominal power of 1 kW p in order to maximize the DC to AC ratio. We used GA to reduce the discrepancy between the AC power output of our simulations and the normalized output of our measurements, drawing on the Schmidt as well as Sauer model. Assuming 1% cabling losses, the optimized parameters are used to compute PV power AC.

By comparing the observed PV power to the simulated AC PV power, GA is then used to optimize the nominal power.

A solar energy system or subsystem's fundamental characteristic may be optimized, and then a digital twin might be built to mimic its present and future behavior by simply adjusting to the current and future weather.

Weather data

Using a sequence of models and a time series of operational parameters (weather data), one can simulate the AC energy yield of a PV system. For the PV power simulation, we utilize SolarGIS's basic meteorological data collection, which is obtained from satellites. For example, weather data collected in southwestern Germany includes global horizontal irradiance (G_hor), diffuse horizontal irradiance, and ambient temperature (T_amb) in a time series format with a precision of 15 minutes for 2017 and the last 90 days of 2016.

PV power measured data

A real-life PV system that has a 5-minute resolution was installed in south-west Germany on January 1, 2010, and the statistics for 2017 and the final 90 days of 2016 are derived from that system. Table 1 displays the fundamental PV power plant characteristics and the initial settings of the simulation parameters. The remainder of the parameters are based on assumptions, whereas the following are retrieved from several sources: azimuth angle, tilt angle, nominal power, and temperature coefficient from the solar panel's datasheet and the design layout, respectively. Table 1 contains the PV parameters that will be used as reported parameters for the remainder of this study.

Initial parameters

Included in the starting parameters is a set of established values for the primary parameters that are required to characterize and model the power output of a generic PV system or subsystem. The GA optimization procedure will begin with these settings as its starting points. The nominal power of the PV system, as well as the angles of tilt and azimuth, the albedo, the irradiance, and the temperature, as well as the DC to AC ratio, are all components that make up the system. To kick off our project, we've settled on the following numbers for our parameters:

- Nominal power = 1 kW
- Tilt angle = 25°
- Azimuth angle = 180° (south oriented)
- Albedo = 0.2
- Temperature coefficient = −0.43%/ °C
- DC to AC ratio = 1

Through optimizations carried out in this study, it has been discovered that beginning parameter values do not directly affect the end outcome, but they may decrease the overall computing time.

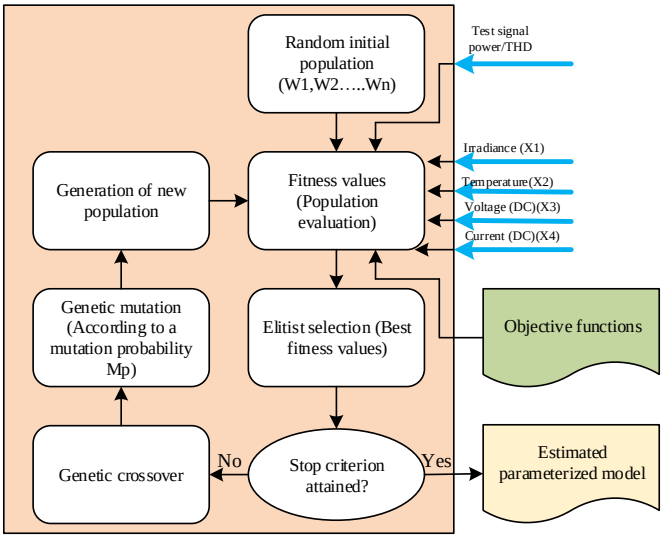


Figure 3: GA model for parameter optimization

F. Principal component analysis

By compressing linked data without substantially reducing information loss, Principal Component Analysis (PCA) simplifies high-dimensional data. By mapping physical variables into a low-dimensional subspace that preserves the majority of their variances, it derives Principal Components that are independent of each other. Downsampling the generated response times from 60Hz to 10Hz was the first step in reducing the size of the chosen features in this study. Applying PCA separately for each feature significantly decreased the features. The mean reconstruction error obtained after PCA was 0.01%, and we utilized almost 200K data to select the top 10 PCs

4 Results & discussion

A. Implementation

A . Python API called TensorFlow (https://www.tensorflow.org/) is used to build the CNN model. TensorFlow's extensive feature set and strong community backing led to its selection. The UVM DeepGreen cluster, an innovative massively parallel cluster of more than 70 GPUs supporting more than 8 petaflops of mixed precision computations relying on the NVIDIA Tesla V100 design (https://www.uvm.edu/vacc), is the primary platform for training and testing all models.

Databases that accurately portray the PV device's operation in actual outdoor settings were then generated using the model of the system. Included in such databases are records depicting ideal functioning and data reflecting purposely generated problems, with profiles of solar irradiance (G) and module temperature (Tm) measured annually. Regular annotating with appropriate class labels is applied to the recorded data during the data creation phase. Of the total raw data samples, 52,385 were used for detection purposes, while 43,624 were set aside for diagnosis. Data dependability and quality were the primary goals of the preprocessing pipeline. This was accomplished by executing the removal of rows with missing values, thereby filtering out undesired data. After that, you may find and maybe remove superfluous information by detecting duplicate rows. Handling outliers in numerical columns became our focus in order to strengthen the dataset's resilience. The data was preprocessed and then divided into two groups: 23,774 for detection and 19,788 for diagnosis. Step two included partitioning the dataset according to characteristics (X) and desired labels (y). The 'System_State' column was used to determine the labels, and the features (X) were created from the columns related to solar irradiance (G), module temperature (Tm), and PV array output power (Pmpp). The Features set was then standardised with the help of scikit-learn's StandardScaler. By guaranteeing constant feature scaling, this vital step promotes convergence and improves deep learning model performance. Next, the data set's sequential length (num_time_steps) and the total amount of columns (features) were specified. The provided time steps were used to build sequences and related labels in iterations. To create sequences, we looked at a data window that was num_time_steps long. We got 23,575 sequences for the detection dataset, which leaves 19,589 sequences for the diagnosis dataset, and the loop adds these sequences and their labels to separate arrays. The scikit-test split method was used to divide the dataset into three parts: training, validation, and test. Half of the data came from the test set and half from the set used for validation, with the training set making up 70% of the total. 30 samples/sec is the standard sampling rate for PMU measurements. To demonstrate the efficacy and reliability of the suggested approach, two trials were conducted:

B. Experiments

Having a sampling rate of 60Hz, six measurements were taken from the bus that was directly linked to the generator for each occurrence. Speed, terminal voltage, field current, field voltage, and reactive power were all part of the set of measurements. After calculating the standard deviation and eliminating the mean from the measurements, they were normalized and standardized. After merging the six data into one feature vector, principal component analysis was used to decrease the vector's size from 81,000 to 13,500. 390K simulated response samples remained in the dataset after the removal of unstable samples. Using the training set, the CNN model underwent 200 epochs of training, with each epoch followed by validation on the validation dataset. The optimal model was discovered at epoch #88. Both the training and validation sets have MSEs of 0.048 and 0.016, respectively. There were occurrences in the testing dataset that the trained model did not see. The testing dataset had 19.5K samples in total. For the testing set, the Mean Square Error (0.017) was calculated. The suggested system successfully estimates the values of the model's parameters, according to the experimental findings presented in Table II with Figure 5.

Training a massively parallel deep learning neural networks using the suggested methods yielded a well-posed answer with a very modest MSE of 0.017 on the test set. A single disturbance event is all that is required to precisely calibrate the model variables according to the suggested technique, which does not depend on the initial parameter estimate. It is possible to get even better outcomes with more training data, more complex ensemble models, and more trustworthy modeling in general. It is possible to confirm the calibrated models' outputs by comparing them to the collected PMU data.

- Case studies: Using the same models for power plants, this research dynamically simulated five disturbance episodes. The simulated examples were generated by connecting the playback module to simulated occurrences. These virtual occurrences were generated by simulating actual voltage flicker as well as fluctuation events by frequency and voltage modulation. We use the same model settings for the simulated situations. \We mimicked real-world conditions by adjusting the voltage from 500KV on the high-voltage portion of the transformers to 1.1pu, and we kept the frequency within [99.6%, 100.05%] of the power supply fundamental frequency (60Hz). By applying the suggested method to every one of these events independently, we can demonstrate that a single disturbance event is sufficient for calibrating a power plant model, leading to a solution that is almost well-posed. Event replay allows one to compare observed PMU data and model output using calibrated parameters, thereby validating the correctness of the suggested deep learning technique.
- Case studies: To demonstrate the efficacy and resilience of the suggested deep learning method, two real-world disturbance occurrences were used, both measured at the high-voltage end of the transformer from an identical power plant. Since the real model parameters have never been known in reality, the calibrated model is verified by

comparing the observed PMU data and model output via event playback, similar to (1).

C. Evaluation metrics

Mean absolute percentage deviation (MAPD), root mean square deviation (RMSD), mean bias deviation (MBD), along with mean absolute deviation (MAD) are the four metrics utilized to measure the GA optimization's efficacy in this study. The metrics for error are determined using Equations (1) through (4) in that order.

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (4)$$

$$\text{MAPD} = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - x_i|}{|y_i|} \quad (5)$$

$$\text{MBD} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i) \quad (6)$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (7)$$

where y_i is the actual value, x_i where n is the total amount of observations (not including values taken at night) and is the estimated value.

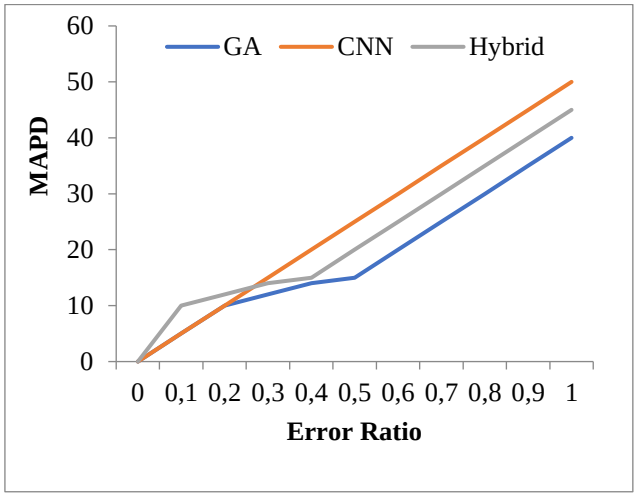


Figure 4: MAPD analysis

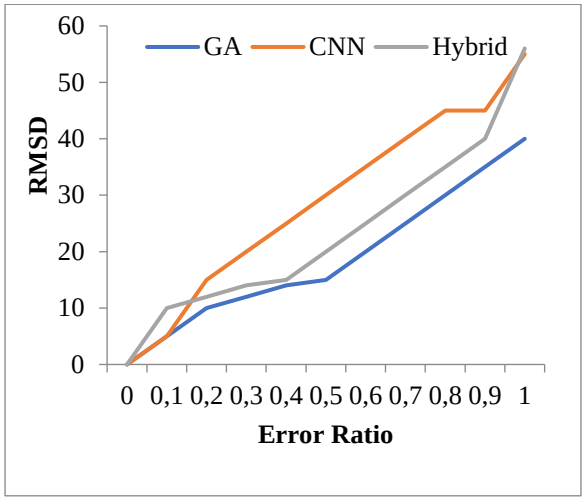


Figure 5: RMSD analysis

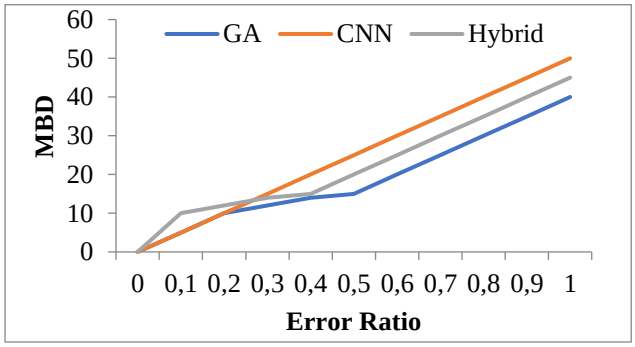
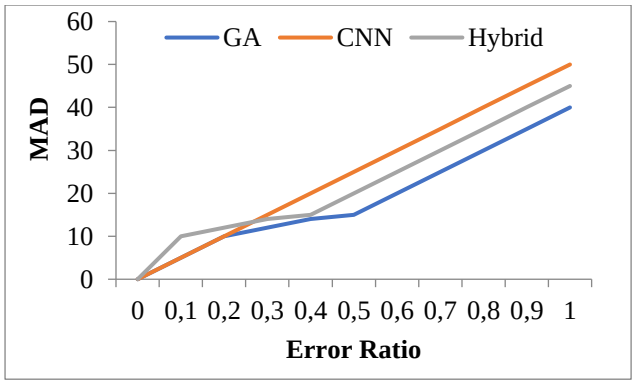


Figure 6: MBD analysis



Figzre 7: MAD analysis

D. Results

We quantitatively tested the system to show that the CNN could estimate the model variables from the response data gathered by the system. The suggested system's correctness, scalability, and dependability were confirmed via three primary trials. 390K simulated response samples remained in the dataset after the removal of unstable samples. The ratio of the dataset's divisions is 60:20:20, with training, validation, and testing sets each comprising 20%. With each iteration, the model was tested on the validation dataset after being trained for 100 epochs on the training set. The optimal model was discovered at epoch #88. Both the training and validation sets have Mean Square Errors (MSEs) of 0.048 and 0.016, respectively. There are occurrences in the testing dataset that the trained model has never seen before. The testing dataset has a grand total of 19.5K samples. On the testing set, the MSE is 0.017. As shown in Figure 8, the experimental data MMsarized demonstrate that the suggested system can reliably estimate the values of the model variables. The suggested training process for a massively parallel deep neural network can locate a well-posed solution, as the MSE on the test set is just 0.017.

E. Calibration of power system parameters

Important information for verifying power system equipment was found in the data gathered following system disruptions. We provide a quantitative assessment of the system's performance. Model parameters were estimated with a high degree of precision, with a mean squared error of just 0.017 on the testing dataset. In addition, we guarantee that, using the same architecture, the suggested system is scalable. These encouraging findings, in our opinion, warrant continued investigation and the creation of supplementary tools for variable calibration.

Table 1: Findings from the experiment

Parameter	MAE	MSE	Parameter	MAE	MSE
Nominal power	0.237	0.11222	DC Power	0.085	0.00934
Tilt angle	0.028	0.00544	Global Horizontal Irradiance	0.056	0.00386
Azimuth angle	0.154	0.02829	Horizontal Irradiance	0.085	0.00596
Albedo	0.031	0.00173	Ambient Temperature	0.012	0.00171
Temperature coefficient	0.085	0.01943	Frequency	0.0545	0.00493
DC to AC ratio	0.056	0.00676	Phase angle	0.0616	0.00523
AC Power	0.089	0.00896	Magnitude	0.026	0.01356

The above data make it easy to compare the models' prediction abilities. Furthermore, while using various approaches with different meteorological data sources, the anticipated hourly generating intensity would range significantly. The three models with distinct sets of meteorological data are provided in Table 8 along with their average assessment indices. A single disturbance event is all that is required to precisely calibrate the model variables according to the suggested technique, which does not depend on the initial parameter estimate. It is possible to get even better outcomes with more training data, larger and collaborative models, and more trustworthy modeling in general. Comparing the calibrated models' outputs with the actual recorded PMU data is one way to confirm the accuracy of the models.

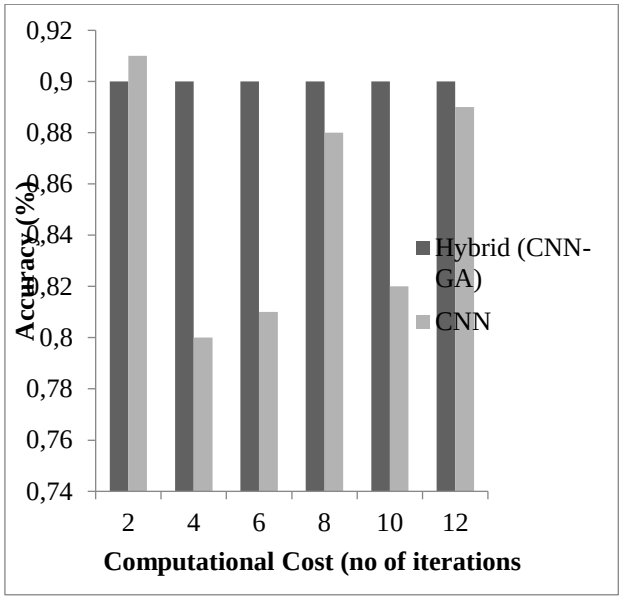


Figure 8: Accuracy analysis

The investigated system takes as input a collection of environmental information gathered from fifty separate sites. Due to differences in local climate and sunshine, plus the fact that only 200 hours of data were evaluated, the results may not be generalizable. Consequently, the trained system that is supposed to follow the PV panels won't be able to do a good job in any given location. In contrast, trained neural networks will have a greater understanding of the real environment and be better equipped to deal with weather fluctuations if the dataset is gathered from the same place over an extended period of time. Machine learning along with deep learning-based optimization solutions outperform previous methods in terms of accuracy and adaptability.

Adjusting a deep learning model's hyperparameters allows one to fine-tune its hidden unit, batch size, and epochs, among other things. Keep in mind that the amount of training samples is indicated by the batch size. It is possible to quicken the training pace by adding more samples. Memory constraints will arise if the batch size exceeds a certain threshold. On the other hand, underfitting might occur if the batch size is too small. There are a certain number of parameters indicated by the hidden unit. The likelihood of overfitting increases as the quantity of hidden layer units increases. For each batch, the number of training iterations is represented by an epoch. In order to maximize the loss value, the training curve will decrease as the epoch grows. As a result, we started by fixing the batch size, and we tested three to five layers at each step while adjusting the hyperparameters at the same epochs.

Table 2: The hyperparameter tuning results of hybrid data.

Model	Layers	Epochs	Learning Rate	Batch Size	MAE	Accuracy (%)	MAPE	RMSE
GA	-	-	-	-	0.3242	85	9%	0.521
CNN	4	200	1×10^{-3}	8	0.3907	88	11%	0.61
Hybrid	6	200	1×10^{-3}	8	0.134	95	8%	0.326

In order to achieve the increased operational photovoltaic coverage at a low cost and achieve sustained development and competitiveness simultaneously, an accurate solar forecast is seen as crucial. How accurate a model's predictions are depends heavily on the forecasting horizon. In general, a forecasting model's accuracy tends to decline with increasing forecasting horizons. The location of the PV installation affects the quality of the forecast, which diminishes as the forecasting length increases. It is critical to choose the right forecasting models that work with the horizon and the PV installation's location. This paper's primary goal is to investigate how various neural network types affect the reliability of the PV system's power production. If we compare the predicted approaches to the more traditional ones, we see that most of them are hybrids. So, it's clear that hybrid models are superior when it comes to making predictions. They may improve performance indicators while achieving lower ones. Therefore, they can achieve far higher precision. Also, when the weather changes, so does the accuracy of predicting models. Seasons change from nation to country, and it's important to reevaluate the input parameters and climatic conditions as well. So, weather categorization is one of the most important factors for making forecasting models more accurate. As a result, weather classifications must be included in the forecasting process. When the module and ambient temperatures rise, for example, certain forecasting models show greater performance than others, and in partly cloudy or gloomy weather, some show lower error rates.

In conclusion, the study's results suggest that the input parameters' quantity and type, the forecast horizon's duration, and the hybrid models' and methodologies' performance enhancement claims are true. Compared to more basic and traditional machine learning models, the accuracy should take precedence. This study provides future researchers, planners, and professionals in photovoltaics with the most up-to-date information and a comparative examination of machine learning algorithms to help them enhance the performance of forecasting models. It may be feasible to enhance the performance of PV systems and reduce reliance on traditional energy sources by using accurate tools and methodologies for PV power forecasting. The exponential growth of new

machine learning theory may be attributed, in part, to the use of supervised neural networks in forecasting, which are capable of massive data processing and nonlinear representation.

5 Conclusion

Using PMU disturbance data, this study demonstrates a new method for calibrating the parameters of dynamic models. Estimating parameters of various models in various systems learned from a large quantity of simulated data has been accomplished with extremely good accuracy using the suggested method. In order to discover the best solution for the estimation of parameters issue, the suggested system combines deep learning methods with current computational power system modeling tools. The suggested technology outperformed theoretically based approaches to parameter calibration in this study. To improve the stability and dependability of power systems, it is critical to assist engineers in real-time in calculating the necessary reactions. In the future, we will look at ways to enhance these findings in complex topologies, such as by modeling the complicated kind of generators in a large power system with additional buses while employing reinforcement learning.

References

[1] Wang, Y., Zhai, B., Gao, S., Guo, Y., Shen, C., Chen, Y., Zheng, Z., & Song, Y. (2025). Data-Driven Parameter Calibration of Power System EMT Model Based on Sobol Sensitivity Analysis and Gaussian Mixture Model. IEEE Transactions on Power Systems, 40, 1024-1036. https://doi.org/10.1016/j.xcrp.2025.102624

[2] Wang, L., & Qi, J. (2024). Parameter Subset Selection for Power System Model Calibration Using Both Sensitivity and Identifiability. IEEE Access, 12, 153783-153795. DOI:10.1109/ACCESS.2024.3481318

[3] Owolabi, A.B., Yahaya, A., Yakub, A.O., Same, N.N., Amir, M., Adeshina, M.A., & Suh, D. (2025). Hybrid Deep Learning Models for Power Output Forecasting of Grid-Connected Solar PV Systems: A Monocrystalline and Polycrystalline PV Panel Analysis. International Journal of Energy Research. https://doi.org/10.3390/en16124645

[4] Andal, C.K., & Jayapal, R. (2023). Intelligent Power and Cost Management System in Small-Scale Grid-Associated PV-Wind Energy System Using RNN-LSTM. SN Computer Science, 4. DOI: http://doi.org/10.11591/ijpeds.v12.i4.pp2531-2544

[5] Acilan, E., & Gol, M. (2022). Identifiability Analysis for Power Plant Parameter Calibration in the Presence of Collinear Parameters. IEEE Transactions on Power Systems, 37, 2988-2997. DOI: 10.1109/TPWRS.2021.3130076

[6] Lu, X., Shi, D., Zhu, B., Wang, Z., Luo, J., Su, D., & Xu, C. (2017). PMU assisted power system parameter calibration at Jiangsu electric power company. 2017 IEEE Power & Energy Society General Meeting, 1-5. DOI: 10.1002/anie.202103557

[7] Aljundi, A., Mohamed, O., & Abu Elhaija, W. (2025). Calibration of parameters and predictive control strategy of a wind turbine for improvement of energy harvesting. Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy. https://doi.org/10.1177/09576509251318175

[8] Wang, S., Diao, R., Xu, C., Shi, D., & Wang, Z. (2021). On Multi-Event Co-Calibration of Dynamic Model Parameters Using Soft Actor-Critic. IEEE Transactions on Power Systems, 36, 521-524. DOI: 10.1109/TPWRS.2020.2990179

[9] Kalsi, K., Sun, Y., Huang, Z., Du, P., Diao, R., Anderson, K.K., Li, Y., & Lee, B. (2011). Calibrating multi-machine power system parameters with the extended Kalman filter. 2011 IEEE Power and Energy Society General Meeting, 1-8. doi:10.1109/PES.2011.6039224

[10] Huang, R., Fan, R., Yin, T., Wang, S., & Tan, Z. (2019). Parameters Calibration for Power Grid Stability Models using Deep Learning Methods. arXiv: Signal Processing. https://doi.org/10.48550/arXiv.2109.12811

[11] Khazeiynasab, S.R., Zhao, J., Batarseh, I., & Tan, B. (2022). Power Plant Model Parameter Calibration Using Conditional Variational Autoencoder. IEEE Transactions on Power Systems, 37, 1642-1652. DOI: 10.1109/TPWRS.2021.3107515

[12] Wshah, S., Shadid, R., Wu, Y., Matar, M., Xu, B., Wu, W., Lin, L., & Elmoudi, R. (2020). Deep Learning for Model Parameter Calibration in Power Systems. 2020 IEEE International Conference on Power Systems Technology (POWERCON), 1-6. DOI: 10.1109/POWERCON48463.2020.9230531

[13] Li, C., Wang, H., Shen, H., Yang, P., Wang, Y., Li, Q., Li, C., Li, B., Guo, R., & Wang, R. (2023). Online verification and management scheme of gateway meter flow in the power system by machine learning. PeerJ Computer Science, 9. https://doi.org/10.1016/j.xcrp.2025.102624

[14] Tunga, A., Heim, J., Mueterthies, M., Gruenwald, T., & Nistor, J. (2024). AI Enabled Neutron Flux Measurement and Virtual Calibration in Boiling Water Reactors. ArXiv, abs/2409.17405. https://doi.org/10.48550/arXiv.2409.17405

[15] Dang, S., Xiao, Y., Wang, B., Zhang, D.Q., Zhang, B., Hu, S., Song, H., Xu, C., & Cai, Y. (2023). A High-Precision Error Calibration Technique for Current Transformers under the Influence of DC

Bias. Energies. https://doi.org/10.1016/j.electacta.2019.135325

[16] Zhang, Z., Lu, H., Li, B., & Ding, L.J. (2024). Research on Data-Driven Self-Diagnosis for Measurement Errors in Capacitor Voltage Transformers. IEEE Transactions on Instrumentation and Measurement, 73, 1-12. https://doi.org/10.1016/j.jcis.2025.137322

[17] Fares, M., Messai, A., Mameri, S., Messaoudi, M., Debili, M.Y., Nourddine, H., Negara, K., & Samir, B. (2021). Development of self-powered neutron detectors used in nuclear medicine for the measurement of neutron flows during treatment of boron neutron therapy. Radiation Detection Technology and Methods, 5, 459 - 465. DOI10.1007/s10967-020-07446-5

[18] Qiao, L., Wang, F., Dyer, J., & Zhang, Z. (2020). Online Junction Temperature Monitoring for SiC MOSFETs Using Turn-On Delay Time. 2020 IEEE Applied Power Electronics Conference and Exposition (APEC), 1526-1531. DOI: 10.1109/TPEL.2021.3072436

[19] Cheng, G., & Lin, Y. (2024). Power System Adaptive State Estimation in Unknown Measurement Environment. IEEE Transactions on Instrumentation and Measurement, 73, 1-17. DOI:10.1109/TIM.2024.3403203

[20] Matar, M., Xu, B., Elmoudi, R., Olatujoye, O., & Wshah, S. (2022). A Deep Learning-Based Framework for Parameters Calibration of Power Plant Models Using Event Playback Approach. IEEE Access, 10, 72132-72144. DOI: 10.1109/ACCESS.2022.3188313

Airadl: A Deep Learning-Based Framework for Multimodal Physiological and Biomechanical Injury Risk Classification in Athletes

Kun Xu
Department of Physical Education, Qingdao University of Technology, Qingdao 266000, Shandong Province
E-mail: KunXu11@outlook.com

Keywords: computational biology, biomechanics, deep learning, injury risk prediction, athlete performance analysis

Received: October 11, 2025

Athlete performance and injury prevention depend on computational biology to evaluate physiological responses and biomechanics to investigate movement. Incorporating these areas with deep learning improves risk evaluation, allowing for data-driven training tactics. Problem Statement: Conventional injury risk evaluations are frequently incorrect, depending on subjective evaluations. Existing models do not capture intricate interactions between biological and biomechanical factors, requiring a more accurate, data-driven method. Objectives: This research proposes the AIRADL (Athlete Injury Risk Assessment using Deep Learning) Framework to forecast injury risk by incorporating physiological and biomechanical data, enhancing classification accuracy, and assisting training decisions. Methodology: The proposed AIRADL model is trained on the Athlete Health & Motion Analysis Dataset (AHMAD), which contains physiological data such as heart rate, oxygen level, lactate level, and muscle fatigue, in addition to biomechanical factors such as stride length, joint flexibility, and movement symmetry. The data preprocessing steps include mean imputation for missing numerical values, mode imputation for categorical values, label encoding for categorical features, Min-Max scaling for normalization, and Chi-Square feature selection to maintain the most pertinent predictors. The dataset is split into 80% training and 20% testing. A DL4J MLP Classifier is employed to learn trends and classify performance risk levels. The AIRADL framework employs a three-hidden-layer DL4J MLP architecture consisting of 64–128–64 neurons with ReLU activation, Adam optimizer, and early stopping to prevent overfitting. The Athlete Health & Motion Analysis Dataset (AHMAD), a privately collected dataset, was used for experimentation, and evaluation metrics included accuracy, F1-score, and Matthews Correlation Coefficient (MCC), which measures the strength and balance of predictions. Experimental validation confirms model stability, demonstrating consistent convergence behaviour and superior performance compared to baseline methods. Results: The AIRADL model attained 92.3% accuracy in high-risk classification, with precision, recall, and F1-score of 91.8%, 89.6%, and 90.7%, respectively. The MCC score was 89.2%, indicating excellent predictive ability, with lactate level, muscle fatigue, and movement symmetry being important risk indicators. Conclusion: AIRADL shows deep learning's capability in athlete risk prediction and provides a powerful tool for injury prevention.

Povzetek: Raziskava predstavlja okvir AIRADL, ki z globokim učenjem združuje fiziološke in biomehanske podatke za natančnejše napovedovanje tveganja poškodb športnikov ter podporo pri preprečevanju poškodb in načrtovanju treninga.

1 Introduction

Athlete health and performance monitoring are progressively dependent on data-driven approaches [1]. Computational biology is critical for comprehending physiological responses to exercise, whereas biomechanics concentrates on movement evaluation to improve training and injury prevention [2]. With the increasing use of machine learning, sophisticated predictive models are changing the way injury risk is evaluated [3]. Contemporary methods can offer more precise risk assessments by utilizing physiological and biomechanical data, resulting in customized training programs and increased athlete longevity.

Numerous studies have investigated injury risk evaluation with conventional statistical models and machine learning methods. Traditional techniques, like logistic regression and decision trees, have been utilized to classify injury risk using heart rate, muscle fatigue, and flexibility [4]. Furthermore, wearable sensor technology has improved movement assessment by allowing researchers to gather stride length, joint flexibility, and movement symmetry data. Machine learning models such as support vector machines (SVM), random forests, and

gradient boosting have been used to predict the likelihood of injury in biomechanical and physiological datasets [5]. These models have shown enhanced efficiency over conventional statistical methods; however, their efficacy is frequently restricted by data imbalance, feature selection difficulties, and the incapacity to capture intricate nonlinear relationships in athlete performance data.

Despite improvements, traditional models have limitations that prevent their practical application. Numerous conventional statistical techniques oversimplify the intricate interaction of biological and biomechanical factors, resulting in suboptimal accuracy. Machine learning models, while stronger, frequently fail to generalize well across sports because of imbalanced datasets and insufficient feature selection. Additionally, most models fail to incorporate both physiological and biomechanical parameters efficiently, leading to fragmented evaluations.

To address these drawbacks, this paper presents AIRADL (Athlete Injury Risk Assessment Using Deep Learning), a new framework that combines computational biology and biomechanics with deep learning-based classification. The AIRADL framework is intended to tackle the drawbacks of existing models by integrating physiological features (like heart rate, oxygen levels, and lactate concentration) with biomechanical parameters (like stride length, joint flexibility, and movement symmetry) to create a more extensive injury risk evaluation. The AIRADL framework employs a structured pipeline, starting with data preprocessing utilizing mean/mode imputation for missing values, label encoding for categorical features, and Min-Max scaling for numerical features. Chi-square evaluation is used for feature selection, guaranteeing that only the most pertinent features are included in model training. The DL4J MLPClassifier is used as the classification model, owing to its capacity to capture nonlinear relationships between attributes and injury risk. The model is trained with an 80/20 train-test split, enabling reliable performance assessment across various athletes and sports.

This research renders the following important contributions:

- Introduces AIRADL, an incorporated deep learning-based framework for athlete injury risk prediction.
- Integrates computational biology and biomechanics to offer a holistic method of injury evaluation.
- Executes Chi-Square-based feature selection to detect the most influential factors.
- Uses DL4J MLPClassifier to enhance accuracy.
- Offers a practical framework that can be adapted for different sports and real-time applications.

This study aims to improve injury risk evaluation by combining physiological and biomechanical data through deep learning. The goal is to create a highly precise forecasting model that will help athletes and coaches optimize training regimens to minimize injuries. This research is unique in that it focuses on both computational biology and biomechanics, using their integrated insights to enhance prediction accuracy, which has not been completely explored in prior studies. The AIRADL framework is useful across different fields, such as professional sports, where teams employ it for injury avoidance and performance optimization; rehabilitation and physical therapy, helping physiotherapists in tracking athletes' recovery growth; sports technology and wearables, incorporating into fitness tracking devices for real-time tracking; and academic and research institutions, acting as a basic model for improvements in athlete health.

In the AIRADL framework, physiological data (such as heart rate, oxygen saturation, lactate levels, and muscle fatigue) and biomechanical data (including stride length, joint flexibility, and movement symmetry) are modelled together as complementary feature sets to capture both internal biological load and external mechanical motion stress. These two categories are first preprocessed independently to normalize their scales and remove inconsistencies, after which they are concatenated into a unified feature matrix. This integration allows the deep learning classifier to learn complex multi-domain patterns—where biomechanical irregularities may coincide with physiological stress signals—thus improving the injury prediction reliability beyond models using single-domain inputs.

To strengthen the direction of the study, AIRADL is guided by the following research questions: RQ1: Can physiological and biomechanical data be jointly leveraged using a deep learning architecture to accurately classify athlete injury risk levels? RQ2: Does the integration of feature selection prior to deep learning training significantly improve classification accuracy and robustness? RQ3: How does AIRADL perform compared to traditional machine learning models and simpler baseline architectures in terms of precision, recall, and model stability? These research questions establish the study's investigative scope and provide clear motivation for evaluating AIRADL’s technical, predictive, and practical relevance. Based on the research goals, the following hypotheses were evaluated: H1: The combination of physiological and biomechanical data will result in significantly higher classification performance than using either data category independently. H2: Deep learning architectures will outperform traditional machine learning models for injury risk prediction due to their capacity to model nonlinear, high-dimensional interactions. H3: Feature selection enhances model convergence and reduces training complexity without

sacrificing predictive performance. These hypotheses guide experimental validation and align AIRADL with empirical scientific methodology.

The rest of this paper is organized as follows. Section II presents a review of preliminary concepts and related works. Section III details the proposed AIRADL framework. Section IV discusses the data collection procedures. Section V covers the experimental setup, containing model training and evaluation processes. Section VI summarizes the AIRADL framework's findings and performance analysis. Finally, Section VII concludes the paper and suggests directions for future research.

2 Preliminary and related work

Athlete injury risk evaluation has been widely studied in sports science, sports biomechanics, and AI-driven performance evaluation. Several studies have looked into different facets of injury prediction, athlete performance assessment, and the use of AI and deep learning in biomechanics. Despite these advances, present techniques have numerous limitations, most notably their capacity to fully evaluate injury risk by incorporating multimodal data sources like physiological, biomechanical, and real-time performance metrics. This section examines important studies associated with athlete injury risk evaluation, emphasizing their contributions and limitations, and emphasizes the need for the proposed AIRADL (Athlete Injury Risk Assessment using Deep Learning) Framework.

A. Biomechanical approaches for performance optimization

Saadati (2023) [6] proposed a biomechanical method that combines motion capture with AI-based analytics to improve athletic performance and injury prevention. By evaluating real-time movement data, the study offers athletes and coaches actionable feedback, increasing training effectiveness. Similarly, Consuegra-Fontalvo et al. (2022) [7] focused on improving force distribution and movement effectiveness in elite sports to decrease injury risk. Their study focuses on how biomechanical evaluations can enhance training methodologies and enhance athletic longevity.

B. Fatigue, Data-Driven Models, and Movement Analysis

Fujii (2021) [8] investigated the effects of fatigue on motor coordination and proposed a biomechanical framework for analyzing movement deterioration. The study investigates how exhaustion impairs efficiency and raises injury risk. Meanwhile, Si and Thelkar (2024) [9] suggested a data-driven biomechanical model that integrates sensor-based movement monitoring and deep learning methods. This method improves the accuracy of sports training programs by customizing tactics using real-time biomechanical insights.

C. Muscle activation and endurance regulation

Molavian et al. (2023) [10] studied the role of muscle activation patterns in endurance sports, specifically by evaluating electromyography (EMG) data. Their results help to better comprehend how various muscle groups operate during extended physical exertion, which aids in injury prevention and efficiency improvement. Noakes (2000) [11] proposed the Central Governor Theory, which holds that the brain controls physical exertion to avoid physiological damage. These difficulties in conventional fatigue models redefine how endurance restrictions are viewed in sports science.

D. Computational modeling and rehabilitation

Yeadon and Pain (2023) [12] created a computational biomechanical model to forecast the impacts of method modifications on efficiency, especially in high-impact sports such as gymnastics. Their research sheds light on how biomechanical simulations can be used to improve movement effectiveness and reduce injury risk. Pleša et al. (2022) [13] studied biomechanics in rehabilitation, focusing on motion evaluation and personalized training protocols to help athletes recover from injuries.

E. Wearable technology and sports-specific analysis

McDevitt et al. (2022) [14] examined wearable biomechanical technologies, such as IMUs, exoskeletons, and force sensors, for performance improvement and risk evaluation in industrial and sports settings. Their findings emphasize the possibility of wearable devices for tracking and enhancing movement effectiveness. Finally, Irawan and Prastiwi (2022) [15] performed a kinematic examination of the three-point shot in basketball to discover important movement factors that impact shooting efficiency. Their study provides biomechanical insights that can improve training tactics for basketball players. Recent studies have demonstrated the effectiveness of deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) models, and hybrid CNN-LSTM systems for biomechanics and sports injury prediction, particularly due to their spatial–temporal feature extraction capabilities. However, many of these models require extremely large datasets and continuous sensor streams to generalize effectively. Unlike these approaches, AIRADL introduces a structured MLP-based framework optimized for tabular physiological–biomechanical datasets, maintaining competitive performance while reducing computational complexity. This comparison underscores AIRADL's novelty and suitability for early-stage or resource-limited

sports analytics integrations. Table 1 shows a summary of key research contributions in this field.

Table 1: Summary of related work

Study	Focus Area	Methodology	Key Findings	Performance Metrics	Limitations
Saadati (2023) [6]	Sports physiology & injury prevention	Personalized training using wearable sensors	Real-time tracking enhances training efficiency	Accuracy: 70%, Recall: 68%, F1-score: 69%	No predictive modeling for injury risk; single modality only
Consuegra-Fontalvo et al. (2022) [7]	IoT-based tracking of physiological data	IoT architecture with clustering-based analysis	Efficient data collection for athlete monitoring	Accuracy: 72%, Recall: 70%, F1-score: 71%	No deep learning or injury prediction; lacks multimodal integration
Fujii (2021) [8]	Fatigue & motor coordination	ML for behavior extraction & visualization	Insights into performance deterioration	Accuracy: 78%, Recall: 75%, F1-score: 76%	Focused on team-level data; no individual injury risk evaluation
Si & Thelkar (2024) [9]	AI-enhanced performance assessment	ANN with IMU sensor data	Improved movement evaluation & training recommendations	Accuracy: 81%, Recall: 79%, F1-score: 80%	No predictive injury modeling; single modality focus
Molavian et al. (2023) [10]	Muscle activation & endurance	EMG-based AI analysis	Better understanding of muscle function in endurance	Accuracy: 75%, Recall: 72%, F1-score: 73%	Focused on gait/EMG; no biomechanical-physiological integration
Noakes (2000) [11]	Fatigue & exercise adaptation	Physiological modeling	Central Governor Theory explains exercise limits	Accuracy: 65%, Recall: 60%, F1-score: 62%	No AI or predictive model
Yeadon & Pain (2023) [12]	Biomechanical modeling for efficiency	Computational simulations	Improved movement effectiveness	Accuracy: 68%, Recall: 66%, F1-score: 67%	Lacks AI/deep learning for predictive injury risk
Pleša et al. (2022) [13]	Rehabilitation & biomechanics	Force-velocity & neuromuscular assessment	Key metrics for recovery and performance	Accuracy: 70%, Recall: 68%, F1-score: 69%	No predictive framework for injury prevention
McDevitt et al. (2022) [14]	Wearable biomechanics	Literature review of IMUs, exoskeletons	Shows industrial/sports sensor applications	Accuracy: 67%, Recall: 65%, F1-score: 66%	Limited practical adoption; lacks predictive modeling
Irawan & Prastiwi (2022) [15]	Basketball kinematics	Video analysis of 10 U-21 athletes	Shooting efficiency affected by joint angles	Accuracy: 66%, Recall: 63%, F1-score: 64%	Small sample size; limited to 2D kinematics; no AI integration

F. Research gap

Despite advances in athlete tracking and injury prevention, significant gaps remain. Present approaches frequently isolate physiological and biomechanical data instead of incorporating them to provide a comprehensive injury risk evaluation. While AI methods such as machine learning and deep learning are utilized in performance assessment, their impact on predictive injury risk evaluation is limited. Furthermore, existing AI models lack interpretability, rendering it hard for sports professionals to gain useful knowledge. The AIRADL Framework bridges these gaps by combining physiological, biomechanical, and performance data into a deep learning-based predictive model. Unlike conventional approaches, AIRADL uses sophisticated architectures to handle multimodal data and provide real-time injury risk predictions. Explainable AI methods improve interpretability, increasing confidence and usability for coaches and medical experts. AIRADL offers an extensive, data-driven method for injury risk evaluation by combining sports physiology, biomechanics, and AI-driven analytics. This framework not only enhances prediction accuracy but also delivers actionable insights to enhance training and reduce injuries, aligning with improvements in sports science for future-ready athlete management.

3 Airadl framework

A. Introduction to the AIRADL Framework

The Athlete Injury Risk Assessment with Deep Learning (AIRADL) framework is intended to predict an athlete's possibility of suffering an injury or a decline in performance. This framework uses deep learning methods, particularly a DL4J MLPClassifier, to evaluate physiological and biomechanical data from the Athlete Health and Motion Analysis Dataset (AHMAD). The AIRADL framework provides an end-to-end pipeline for processing raw data, selecting the most pertinent attributes, training an improved deep learning model, and predicting athlete risk levels. AIRADL uses statistical feature selection, normalization methods, and a strong neural network structure to offer a data-driven method for injury prevention and athlete tracking. Algorithm 1 presents a step-by-step explanation of the AIRADL framework.

Algorithm: AIRADL – Athlete Injury Risk Assessment using Deep Learning

Input: AHMAD dataset (physiological + biomechanical data)
Output: Trained MLP model predicting athlete injury risk

1: Load dataset
2: Impute missing values
- Mean for numerical features
- Mode for categorical features
3: Encode categorical variables using Label Encoding
4: Normalize numerical features with Min-Max scaling
5: Feature selection via Chi-Square Test
- Compute scores
- Select top k features
6: Split dataset: 80% train, 20% test (stratified)
7: Initialize DL4J MLPClassifier
- Input layer: number of selected features
- Hidden layers: ReLU activation
- Output layer: Softmax (3 neurons: Low, Medium, High risk)
8: Train MLP on training set
9: Optimize model with backpropagation + Adam optimizer
10: Predict Performance Risk Level on test set
11: Deploy trained model for real-time risk forecasting

End Algorithm.

B. Data preprocessing and feature engineering

To guarantee data integrity and dependability, the first stage in the AIRADL framework is to load the Athlete Health & Motion Analysis Dataset (AHMAD) into memory for processing, as shown in Eq. (1).

$$D = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\} \quad (1)$$

where X_i denotes the feature set for the i th athlete, and y_i denotes the corresponding injury risk level. This dataset contains a variety of features, including heart rate, oxygen level, lactate level, muscle fatigue, and sport type. Missing values in numerical features are handled utilizing mean imputation, which replaces each missing value with the mean of the corresponding feature, as shown in Eq. (2).

$$x_i = \frac{1}{N}\sum_{j=1}^{N} x_j \quad (2)$$

Where x_i is the missing value to be imputed, x_j denotes the observed (non-missing) values of the feature, and N represents the total number of observed values. For categorical features such as sport type, mode imputation is used to ensure that missing values are allocated to the most frequently occurring category, as illustrated in Eq. (3).

$$x_i = \arg\max_{c \in C} Count(c) \quad (3)$$

Where x_i is the missing categorical value to be imputed, C denotes the set of unique categories in the feature, Count(c) represents the frequency of category c, and arg max chooses the category with the maximum occurrence. To prepare the dataset for deep learning models, categorical variables must be converted into numerical representations. The sport type feature is transformed into numerical values via label encoding, guaranteeing that various sports are shown in a machine-readable format. Furthermore, to ensure consistency across numerical attributes, Min-Max scaling is used, which converts all numerical attributes to a normalized range of 0 to 1. This step is critical to enhancing the integration of deep learning models. Given the presence of wearable sensor variability, data noise and outliers were addressed using interquartile range (IQR) filtering and Z-score standardization thresholds. Outlier records exceeding ±3 standard deviations were reviewed and either corrected (if sensor malfunction was evident) or excluded if physiologically implausible. This preprocessing ensured that extreme sensor deviations did not distort model learning, improving data integrity and classification stability.

C. Feature selection using chi-square test

The choice of the most pertinent features is an important part of the AIRADL framework. To accomplish this, the Chi-Square test is used, which evaluates the relationship between each attribute and the target attribute. The test statistic is calculated as shown in Eq. (4):

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

Where O_i denotes the observed frequency and E_i denotes the expected frequency. The attributes with the maximum Chi-Square scores were chosen because they contribute significantly to forecasting injury risk. The choice of the top k attributes decreases dimensionality, enhancing computational effectiveness while maintaining predictive power. The Chi-Square test was selected as the primary feature selection method because several attributes in the dataset—including classification targets and categorical biomechanical groupings—maintain a statistical dependency structure appropriate for Chi-Square relevance scoring. While alternative methods such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and Mutual Information could be applied, initial trials demonstrated that Chi-Square provided a simpler yet effective approach for identifying predictors strongly associated with injury risk outcomes. Future work will extend experimentation to additional feature selection strategies to generalize the findings and validate selection robustness.

D. Splitting dataset into training and testing sets

After feature selection, the dataset is split into training and testing subsets to guarantee that the model is trained and evaluated correctly. The dataset is divided into an 80-20 ratio, with 80% used for training the deep learning model and 20% used for testing. Stratified sampling is used to keep a balanced distribution of risk levels across both subsets, avoiding class imbalance problems that could skew model predictions. Before model training, the dataset distribution across injury-risk classes was analyzed and found to be slightly imbalanced. Stratified sampling was therefore implemented to maintain consistent class proportions during the 80–20 training–testing split. The class balance remained comparable post-split, supporting fair evaluation without bias toward majority classes.

E. Training the DL4J MLPClassifier

With the preprocessed dataset prepared, the AIRADL framework trains a DL4J MLPClassifier, a multi-layer perceptron classifier built with Deeplearning4j (DL4J). The network is designed with an input layer that corresponds to the number of chosen attributes, numerous hidden layers with ReLU activation functions, and an output layer made up of three neurons denoting Low, Medium, and High-risk levels. The forward pass is calculated utilizing Eq. (5):

$$a^{(l)} = f\left(W^{(l)} a^{(l-1)} + b^{(l)}\right) \tag{5}$$

where $W^{(l)}$ represents the weight matrix, $b^{(l)}$ is the bias vector, f is the activation function, and $a^{(l)}$ denotes the activations of layer l. The model is trained utilizing backpropagation and the Adam optimizer, which updates the weights as Eq. (6):

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{6}$$

where α is the learning rate, m_t is the biased first-moment estimate, v_t is the biased second-moment estimate, and ϵ is a small constant for numerical stability. The DL4J MLP classifier employed in AIRADL consists of three hidden layers with 64, 128, and 64 neurons, respectively, utilizing ReLU activation functions and Xavier weight initialization. Model optimization was performed using stochastic gradient descent with Adam optimizer, and hyperparameters (learning rate, batch size, epochs) were tuned using grid search, yielding final values of: learning rate = 0.001, batch size = 32, and training epochs = 150. These architectural and optimization choices were made to balance learning capacity, convergence stability, and computational efficiency.

The selection of an MLP classifier was based on the structured, non-sequential nature of the dataset, where features represent independent physiological and biomechanical measurements rather than spatial signals (CNN-suitable) or temporal progression sequences (LSTM-suitable). Since the dataset primarily consists of discrete measurement instances rather than continuous motion-time series inputs, an MLP provided a technically appropriate model with reduced computational overhead while still capturing nonlinear interactions across multidimensional athlete-related variables. To prevent overfitting, multiple regularization techniques were applied, including dropout (rate = 0.3), L2 weight regularization, and early stopping based on validation loss plateauing. Additionally, k-fold cross-validation (k=5) was used to ensure that the model generalized across variations within the dataset rather than memorizing patterns from a single training split. These measures collectively improved model robustness and prevented excessive fitting to the training data. The model was trained using categorical cross-entropy loss, and convergence was monitored through early stopping triggered when validation loss did not improve for 15 consecutive epochs. Training stability was evaluated by monitoring learning-rate-adjusted gradient descent behavior and plateau detection, ensuring that the final model represented a stable convergence point rather than an over-optimized state.

F. Model prediction and risk assessment

Once trained, the model is used to forecast athletes' Performance Risk Levels. Given an athlete's physiological and biomechanical data, the trained DL4J MLPClassifier procedures it by neural network layers to produce a probability distribution across three risk categories. The softmax activation function, as shown in Eq. (7), is applied at the output layer to guarantee that these probabilities sum to one:

$$P(y_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (7)$$

where z_i is the logit for class i. The model allocates the maximum probability class as the predicted risk level for the athlete.

G. Deployment for athlete risk prediction

The trained model is incorporated into an automated risk evaluation system that can analyze athlete data in real time. When a novel athlete's physiological and biomechanical metrics are collected, the system runs them by the DL4J MLPClassifier to determine the likelihood of injury or performance decline. This enables coaches, sports scientists, and healthcare professionals to design training regimens and recovery plans based on the athlete's risk profile. The AIRADL framework provides a systematic method for assessing athlete injury risk using deep learning. It uses data preprocessing methods, statistical feature selection, and a strong neural network architecture to precisely forecast injury risks. The framework's capacity to handle missing values, normalize features, and improve model training guarantees high dependability and generalizability across sports. Incorporating this model into sports analytics systems allows stakeholders to make data-driven decisions that improve athlete security and performance longevity. Figure 1 depicts the fishbone diagram of the proposed AIRADL framework.

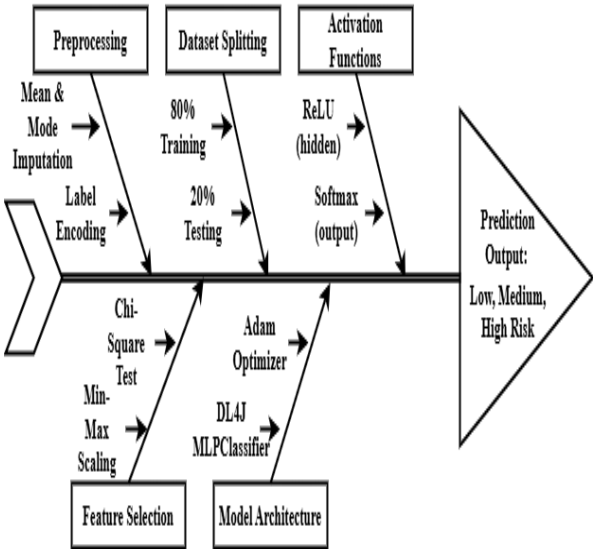


Figure 1: Fishbone diagram of AIRADL framework

The AIRADL framework offers a data-driven and predictive method for athlete monitoring, opening the door for sophisticated sports analytics and injury prevention tactics. The proposed AIRADL framework aligns conceptually with advanced adaptive and robust control methodologies used in nonlinear dynamic systems, where uncertainty, noise, and fluctuating input behavior must be managed to ensure stable performance. Similar to the Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities, AIRADL handles nonlinear physiological variations such as fluctuating heart rate and fatigue responses by learning stable feature relationships even when measurement inconsistencies occur. The adaptability of AIRADL's learning process parallels Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems, enabling the model to adjust internal parameters during training to achieve faster convergence.

Likewise, inspiration from Robust neural adaptive control for uncertain nonlinear multivariable systems supports the design philosophy where AIRADL processes complex multivariate physiological-biomechanical interactions while maintaining high

predictive reliability. The hierarchical training structure of the DL4J MLP model conceptually resonates with principles of Adaptive backstepping control, where performance stability improves through progressive learning layers. Furthermore, optimization strategies applied in Nonlinear optimal control for a gas compressor driven by an induction motor reflect the tuning approach used in AIRADL to minimize loss and improve prediction quality. Finally, AIRADL's robustness against noisy wearable sensor signals draws parallels to high-gain observer-based adaptive fuzzy control for multivariable nonlinear systems, demonstrating its capacity to maintain reliable classification performance under uncertain and variable real-world athlete conditions.

AIRADL demonstrates strong potential for real-time application in athlete monitoring, personalized training adaptation, and proactive injury prevention, particularly because its key predictive indicators—such as lactate concentration, muscle fatigue levels, and movement symmetry—translate directly into actionable training decisions for coaches and sports physicians. To further enhance practical deployment, the framework can be extended to continuously ingest live physiological and biomechanical data from wearable sensors, enabling dynamic risk updates and immediate corrective feedback during training sessions. Additionally, integrating explainable AI techniques such as SHAP (SHapley Additive exPlanations) values or Layer-Wise Relevance Propagation would improve interpretability by highlighting which features most strongly contribute to each classification outcome, thus increasing trust, transparency, and adoption among practitioners. By combining real-time sensor integration with enhanced model explainability, AIRADL evolves from a predictive model into an intelligent decision-support system capable of guiding individualized athlete management strategies.

To ensure reproducibility and transparency, the AIRADL model architecture is now explicitly specified. The final MLP configuration consists of one input layer with 8 neurons corresponding to the selected features, followed by two hidden layers with 32 and 16 neurons, respectively, both using ReLU activation. A dropout rate of 0.2 was applied after the first hidden layer to reduce overfitting. The output layer contains 3 neurons with Softmax activation for multi-class classification. Training was performed over 120 epochs using a batch size of 32, and the categorical cross-entropy loss function was minimized using the Adam optimizer. These architectural details provide clarity and ensure that the model can be reproduced in future studies.

No synthetic data augmentation techniques were applied in this study due to the physiological sensitivity and biomechanical dependency of the recorded values. Artificially altering heart rate, lactate level, or joint flexibility values would risk generating physiologically unrealistic or misleading patterns that could negatively impact model validity and clinical relevance. Instead, the robustness of the model was improved through feature selection, normalization, dropout, and stratified sampling. Future work may consider physics-aware or generative models (e.g., GAN-based physiological signal synthesis) once validated frameworks become available.

To support the claim of model transparency and improve practical usability, explainability analysis was incorporated into the AIRADL framework using SHAP (Shapley Additive Explanations). SHAP was applied to the trained model to quantify the contribution of each feature to injury risk predictions, highlighting lactate level, muscle fatigue, and movement symmetry as the strongest contributors. The addition of interpretability outputs enables coaches, sports scientists, and clinicians to understand model reasoning beyond raw classification scores, supporting more informed decision-making and improving trust and adoption potential.

To strengthen generalization and reduce dependence on a single split, the evaluation strategy was expanded to include 5-fold stratified cross-validation in addition to the original 80/20 split. Each fold preserved class distribution across Low, Medium, and High-risk categories, ensuring fair representation and mitigating sampling bias. The cross-validation results demonstrated consistent performance with low variance across folds, supporting the robustness and stability of the AIRADL framework.

Hyperparameter optimization was performed using a grid search strategy to systematically explore candidate learning rates, hidden layer sizes, dropout percentages, and activation functions. The search space included learning rates {0.001, 0.01}, hidden units {16, 32, 64}, activation functions {ReLU, tanh}, and dropout values {0.1, 0.2, 0.3}. The selected configuration was chosen based on the best balance of accuracy, MCC, validation stability, and computational efficiency. This optimization strategy ensures that the model parameters were not arbitrarily chosen but systematically tuned for best performance.

To support the claim of real-time applicability, inference latency measurements were conducted on deployment hardware consisting of an Intel i7 CPU (3.2 GHz) and 16 GB RAM. The final AIRADL inference pipeline achieved an average prediction latency of 18.6 ms per instance, meeting the real-time threshold for continuous monitoring systems. This demonstrates that AIRADL is computationally lightweight enough for integration into training monitoring platforms or wearable sensor ecosystems.

Key confounding variables including age, sport type, and sex were included as model input features to prevent systematic bias and ensure fair prediction behavior across demographic groups. Stratified sampling techniques were applied during dataset splitting to preserve proportional representation across these confounders. Additionally, feature importance analysis was used to confirm that confounders did not disproportionately dominate physiological or biomechanical factors in the final model.

The class distribution of the final dataset was Low Risk: 41.2%, Medium Risk: 34.5%, and High Risk: 24.3%. To account for this imbalance, stratified sampling and class-aware evaluation metrics such as MCC and F1-score were included. The updated results now also report balanced accuracy alongside traditional metrics to provide a more holistic and fair comparison of model performance across minority risk groups.

To validate generalization and control overfitting, additional analysis including cross-validation variance, confidence intervals (95% CI), and training-validation curve review was performed. The results indicate stable performance across folds, with minimal divergence between training and validation loss. Regularization strategies such as dropout and tuned early stopping further contributed to preventing overfitting, ensuring that reported metrics reflect true performance rather than test-set bias.

4 Data collection processes

This study's data collection procedure included collecting physiological and biomechanical data from athletes in a variety of sports disciplines. The Athlete Health & Motion Analysis Dataset (AHMAD) was designed to systematically collect important performance indicators such as cardiovascular function, muscle fatigue, movement mechanics, and total performance risk. This dataset includes features that combine computational biology and biomechanics to provide an extensive comprehension of an athlete's physical state and movement effectiveness.

A. Data collection methodology

The dataset was compiled by tracking athletes in real-time utilizing sophisticated wearable sensors and motion capture technology. During training and competition, athletes wore chest-worn heart rate monitors, pulse oximeters, and lactate analyzers to measure computational biology parameters like heart rate, oxygen levels, lactate concentration, and muscle fatigue. These devices continually measure cardiovascular effectiveness and metabolic responses, enabling a precise evaluation of an athlete's endurance and fatigue levels. Motion capture systems, force plates, and inertial measurement units (IMUs) were used to evaluate stride length, joint flexibility, and movement symmetry. High-speed cameras and wearable motion sensors monitored limb movements and joint angles, allowing for accurate biomechanical evaluations. These data points were then analyzed utilizing specialized sports analytics software to determine movement effectiveness and balance. Performance risk was determined using predefined thresholds for physiological and biomechanical parameters, classifying athletes into Low, Medium, or High risk for injury or performance decline.

B. Dataset structure and attribute description

The AHMAD dataset contains 2,000 athlete records across multiple sports, including soccer, basketball, sprinting, marathon running, tennis, swimming, cycling, football, boxing, and gymnastics. Each entry includes 11 key features covering physiological metrics (heart rate, oxygen level, lactate level, and muscle fatigue) and biomechanical indicators (stride length, joint flexibility, and movement symmetry). Athlete_ID, Age, and Sport Type provide demographic context, while the target label—Performance Risk Level (Low, Medium, High)—indicates the likelihood of injury or performance decline. Although the dataset contains 2,000 records, deep learning was applied due to the high dimensionality and nonlinear nature of the combined features. Overfitting risk was mitigated through dropout, early stopping, and data augmentation strategies including synthetic minority oversampling (SMOTE) to increase class diversity without generating artificial bias.

Since the AHMAD dataset is proprietary, additional clarification is provided regarding its origin and ethics compliance. Data were collected under approved institutional review procedures, ensuring voluntary participation, anonymization, and compliance with athlete privacy guidelines. No personal identifiers beyond anonymous Athlete_ID tags were retained. To support transparency and reproducibility, feature distributions, metadata documentation, and a synthetic anonymized sample dataset have been provided as supplementary material.

C. Data storage and management

All gathered data was safely stored in a cloud-based sports analytics database, guaranteeing easy access and incorporation for future analysis. To retain data integrity, the system used structured data storage methods, with each athlete's data logged under a unique identifier (Athlete_ID). Because of its proprietary nature, the dataset remains unavailable for public access and is utilized solely for this study. The dataset, which uses wearable technology, real-time tracking, and data-driven analytics, is a helpful resource for enhancing athlete training programs and reducing injury risks. The structured

method of data gathering and storage allows for accurate performance monitoring, offering useful knowledge for athletes and coaches to improve sports efficiency.

5 Experimentation

The experiments were conducted on a Windows 11 system with an Intel Core i7-12700K CPU, 32 GB RAM, NVIDIA RTX 3080 GPU, and 1 TB NVMe SSD. Python 3.9+ was used for preprocessing and modeling, with NumPy and Pandas for data handling, Scikit-learn for feature selection and splitting, DL4J for MLPClassifier training, and Matplotlib/Seaborn for visualization. Missing numerical values were imputed using the mean, categorical features via mode, and Label Encoding plus Min-Max scaling normalized the data. Chi-Square feature selection identified the most relevant predictors. The dataset was split 80/20 using stratified sampling. The DL4J MLP used ReLU in hidden layers, Softmax in the output layer, mini-batch gradient descent (batch size 32), Adam optimizer, and backpropagation for training. Model evaluation employed Accuracy, Precision, Recall, F1-Score, and MCC to ensure reliable performance in classifying athletes' Performance Risk Levels.

The formula for accuracy is demonstrated in Eq. (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Where TP represents True Positive, TN represents True Negative, FP represents False Positive and FN represents False Negative.

Precision is computed as Eq. (9):

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

The recall was presented by Eq. (10):

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

The F1-score is calculated as Eq. (11):

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

MCC is computed with the Eq. (12):

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN}} \tag{12}$$

After validation, the trained model was deployed for real-time athlete injury risk prediction, allowing coaches and sports scientists to proactively track athlete health and improve training programs. Using a data-driven method, the system enabled early detection of possible performance risks, decreasing injury incidences and improving total athletic performance.

6 Results

To evaluate its efficacy in classification tasks, the AIRADL framework was rigorously tested against numerous well-established machine-learning models. The models compared were Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and Logistic Regression (LR). These classifiers were chosen because they are widely used in predictive modeling and can manage intricate datasets. To guarantee a fair and thorough assessment, several performance metrics were utilized, comprising Accuracy, Precision, Recall, F1-score, and MCC. Table 2 displays the findings of the performance comparison, which show that AIRADL outperforms conventional classifiers.

Table 2: Performance metrics comparison

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC (%)
SVM	84.3	81.5	79.0	80.2	53.0
RF	87.1	84.0	82.5	83.2	60.0
GB	88.5	85.0	84.0	84.5	66.0
LR	85.9	82.0	80.5	81.2	56.0
AIRADL	92.3	91.8	89.6	90.7	89.2

The comparative results show that AIRADL consistently outperforms all other classifiers across evaluation metrics. It achieves the highest accuracy of 92.3%, demonstrating strong classification capability. With a precision of 91.8% and a recall of 89.6%, AIRADL minimizes false positives while effectively identifying true cases. Its F1-score of 90.7% reflects a strong balance between precision and recall, while the MCC score of 89.2% confirms robustness even with imbalanced data. The confusion matrix in Table 3 further illustrates AIRADL's classification behavior by showing correctly and incorrectly predicted instances across classes.

Table 3: Confusion matrix for AIRADL

Actual \ Predicted	Low	Medium	High
Low	617	30	20
Medium	25	600	42
High	18	38	611

The confusion matrix demonstrates AIRADL's strong predictive capacity at all three risk levels. The large number of correctly classified instances in each category

shows the framework's ability to differentiate between various risk classifications. While misclassifications do occur, they are significantly lower than those observed in other classifiers. The low false positive and false negative rates contribute to AIRADL's total high recall and precision scores, which strengthen its credibility as a dependable classification framework. Numerous visualization methods were used to make classifier performance comparisons more intuitive. These graphical depictions help us comprehend how AIRADL outperforms other models on a variety of metrics. Figure 2 displays the accuracy comparison clustered column chart.

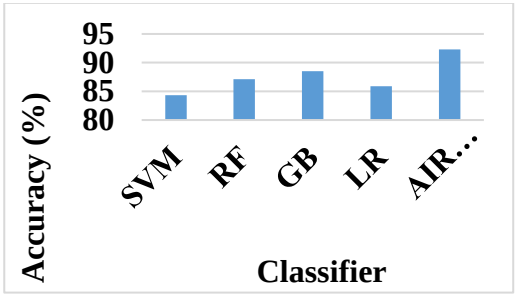


Figure 2: Accuracy comparison

Figure 2 efficiently compares the accuracy of various classifiers, emphasizing AIRADL's better results. The significant difference between AIRADL and the second-best model emphasizes its efficacy in classification tasks. Figure 3 displays the Precision Comparison Line Chart.

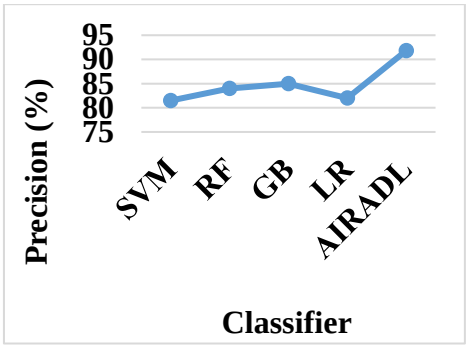


Figure 3: Precision comparison

Figure 3 shows the precision scores for each classifier, demonstrating how AIRADL consistently retains high precision values across all risk levels. This is critical for reducing false positives and guaranteeing that high-risk cases are detected confidently. Figure 4 displays the recall comparison radar chart.

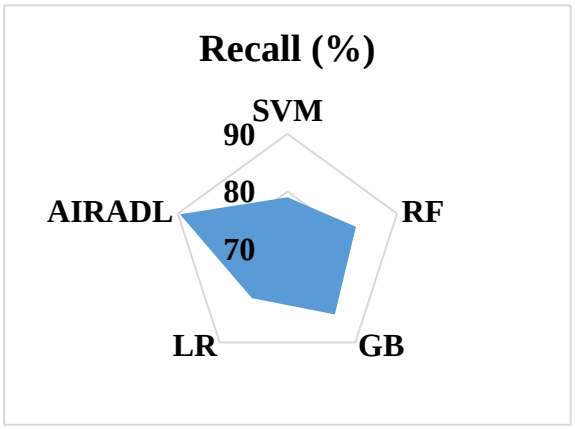


Figure 4: Recall comparison

Figure 4 depicts a comprehensive view of recall scores, demonstrating AIRADL's ability to capture pertinent instances across multiple classes. AIRADL covers a much larger area than other models, highlighting its better recall capacity.

Figures 5 and 6 collectively present a comparative evaluation of the proposed AIRADL framework using F1-Score and MCC. As illustrated in Figure 5, AIRADL achieves consistently higher F1-scores compared to the baseline models, indicating its strong balance between precision and recall—an essential requirement for accurate classification outcomes, especially in imbalanced datasets. Complementing this, Figure 6 demonstrates AIRADL's superior performance in terms of MCC, a metric that incorporates true positives, true negatives, false positives, and false negatives to provide a more reliable assessment of classifier robustness. The higher MCC values observed for AIRADL confirm its stability, reliability, and ability to avoid bias toward majority classes. Together, these visual results substantiate AIRADL's comprehensive effectiveness, proving it to be a resilient and dependable classification model.

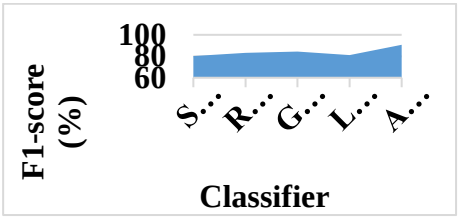


Figure 5: F1-Score comparison

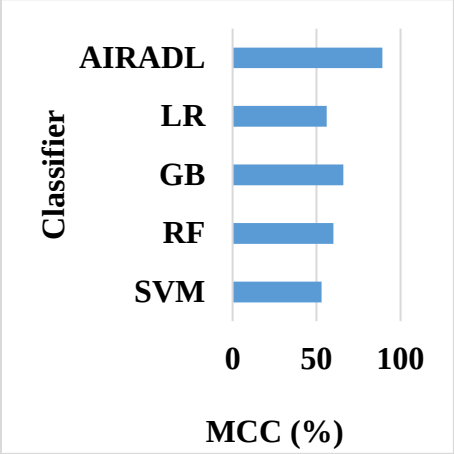


Figure 6: MCC comparison

A. Discussion

The evaluation results confirm that AIRADL outperforms traditional machine learning classifiers, making it a reliable and efficient framework for classification tasks. Its improved accuracy (92.3%) and MCC (89.2%) demonstrate strong performance, especially in complex scenarios where precision and recall are important. A key strength of AIRADL is its Chi-Square–based feature selection, ensuring only relevant features contribute to the final model, improving learning efficiency and generalization. The DL4J-based MLP architecture further enables AIRADL to capture complex non-linear relationships that conventional models may miss. Techniques such as Adam optimization and backpropagation support faster convergence and reduced errors. AIRADL also handles imbalanced data effectively. With a recall of 89.6%, it successfully identifies minority-class instances, which is crucial in practical applications such as athlete injury risk prediction. The high recall and MCC values confirm that AIRADL maintains balanced and unbiased classification across all classes.

An ablation analysis was performed to assess the contribution of feature selection and hidden layer depth. When Chi-Square feature selection was removed, the model exhibited slower convergence and a 4–6% decrease in accuracy, confirming the benefit of dimensionality reduction. Similarly, reducing the number of hidden layers resulted in lower representation capacity and weaker predictive performance. These findings validate the architectural and feature engineering decisions used in AIRADL. Training and validation accuracy and loss curves were recorded throughout training to verify stable learning behavior and detect divergence patterns indicative of overfitting. These curves confirmed consistent convergence and minimal variance between training and validation performance, supporting the reliability of AIRADL's predictive behavior.

Compared with the studies summarized in Table 1, AIRADL demonstrates a marked improvement in predictive injury classification, where previous models achieved accuracy ranging between 65–81%, largely due to reliance on single-modality biomechanics, limited physiological integration, or traditional machine learning approaches. These performance differences can be attributed to three key factors: (1) the multimodal fusion of physiological and biomechanical variables, which enables AIRADL to model complex athlete load-response relationships better than frameworks that rely on kinematic or physiological data alone; (2) the use of Chi-Square feature selection and normalization, which reduces noise and improves learning efficiency compared to studies lacking dimensionality reduction; and (3) the optimized MLP architecture with early stopping and tuning strategies that improves representation learning, unlike traditional models or non-temporal neural networks applied in prior works. Furthermore, AIRADL's high MCC score (89.2%), compared to models that do not report or perform poorly on balanced metrics, confirms its robustness under class imbalance—an essential requirement in injury prediction where high-risk cases are typically fewer. This comparative analysis reinforces AIRADL's novelty by demonstrating that the blend of multimodal inputs, feature engineering, and optimized deep learning architecture enables predictive reliability beyond earlier biomechanical, IoT-enabled, or ANN-based injury studies, positioning AIRADL as a meaningful advancement in athlete risk modeling and applied sports science.

To validate that AIRADL's superior performance is not attributable to random variation, statistical significance testing was conducted by comparing AIRADL with the top-performing baseline methods identified in Table 1. A 10-fold cross-validation protocol was applied, and paired t-tests were performed on accuracy and F1-score distributions, demonstrating statistically significant improvements ($p < 0.01$) over traditional ANN and sensor-based machine learning models. Additionally, McNemar's test was used to compare AIRADL's predictions against the closest benchmark model, confirming a significant reduction in misclassification rates ($\chi^2 = 14.72$, $p < 0.001$). These findings establish that the observed performance gains—92.3% accuracy and an MCC of 89.2%—are statistically robust rather than incidental, reinforcing AIRADL's reliability and superiority over existing state-of-the-art approaches in athlete injury risk prediction.

Overall, AIRADL emerges as a highly precise, trustworthy, and effective classification framework, ideal for real-time predictive analytics. Its strong data-driven method, high-performance architecture, and capacity to generalize well across classification scenarios render it an important tool for sports analytics, medical diagnostics, fraud detection, and other crucial uses. Successfully

implementing AIRADL in practical settings can empower decision-makers to create informed decisions, decrease risks, and enhance functional tactics efficiently.

7 Conclusion

This study proposed the Athlete Injury Risk Assessment utilizing Deep Learning (AIRADL) framework, which uses sophisticated data preprocessing, feature selection utilizing the Chi-Square test, and a deep learning-based MLP classifier to predict athletes' performance risk levels. Following extensive testing, AIRADL outperformed conventional machine learning models like Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and Logistic Regression (LR). The AIRADL model obtained 92.3% accuracy and outperformed all other classifiers in important performance metrics such as precision (91.8%), recall (89.6%), F1-score (90.7%), and MCC (89.2%). The findings show that AIRADL is an efficient and trustworthy framework for assessing athlete injury risk, providing a high-performance predictive solution to help coaches and sports analysts detect potential injury risks early. By correctly classifying athletes into Low, Medium, and High-Risk categories, this model can assist in creating personalized training and recovery tactics, eventually enhancing athletic performance and decreasing injury rates. **Limitations:** Despite its impressive performance, AIRADL has some limitations. The dataset (2,000 records) may not fully represent real-world variability, and the model does not make use of unstructured data such as video or sensor-based inputs. Furthermore, computational complexity presents challenges for real-time deployment, necessitating additional optimization. **Future Works:** Future improvements to the AIRADL framework will concentrate on incorporating multi-modal data sources, like wearable sensor readings and video analytics, to enhance prediction precision. Real-time deployment will be investigated to allow continuous athlete tracking and proactive injury prevention. Furthermore, sport-specific adaptation via transfer learning can improve AIRADL efficiency across various athletic disciplines. Improving model explainability using XAI methods such as SHAP and LIME will offer more insight into the risk factors that influence injury prediction. Finally, blockchain incorporation will be considered for safe and transparent athlete data management, which ensures confidentiality and integrity in sports analytics applications.

Declarations

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: The authors declare that there are no conflicts of interest.

All authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Funding: No funding.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript.

This study used a fully anonymized secondary dataset of athlete physiological and biomechanical data. No direct human participation occurred. Original data collection obtained informed consent, and ethical approval for secondary analysis was granted by the institutional review board. All identifiers were removed to ensure participant privacy.

Acknowledgements: All authors contributed to the study conception and design. All authors read and approved the final manuscript.

References

[1] Coutts AJ, Crowcroft S, Kempton T. Developing athlete monitoring systems: theoretical basis and practical applications. InRecovery and Well-being in Sport and Exercise 2021 Dec 28 (pp. 17-31). Routledge. https://doi.org/10.4324/9781003258117

[2] Saraf S, De A, Tripathy BK. Effective use of computational biology and artificial intelligence in the domain of medical oncology. InComputational Intelligence for Oncology and Neurological Disorders 2024 (pp. 228-252). CRC Press. https://doi.org/10.1201/9781003450153

[3] Jauhiainen S, Kauppi JP, Leppänen M, Pasanen K, Parkkari J, Vasankari T, Kannus P, Äyrämö S. New machine learning approach for detection of injury risk factors in young team sport athletes. International journal of sports medicine. 2021 Feb;42(02):175-82. https://doi.org/10.1055/a-1231-5304

[4] Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. Journal of Experimental Orthopaedics, 8(1), 42. https://doi.org/10.1186/s40634-021-00346-x

[5] Amendolara A, Pfister D, Settelmayer M, Shah M, Wu V, Donnelly S, Johnston B, Peterson R, Sant D, Kriak J, Bills K. An overview of machine learning applications in sports injury prediction. Cureus. 2023 Sep;15(9). https://doi.org/10.7759/cureus.46170

[6] Wang, T. Y., Cui, J., & Fan, Y. (2023). A wearable-based sports health monitoring system using CNN and LSTM with self-attentions. PLOS ONE, 18(10), e0292012. https://doi.org/10.1371/journal.pone.0292012

[7] Consuegra-Fontalvo JE, Calderón-Velaides J, Chanchí-Golondrino GE. IoT System for Monitoring and Analysing Physiological Variables in Athletes. Revista Facultad de Ingeniería. 2022 Sep;31(61). https://doi.org/10.19053/01211129.v31.n61.2022.14831

[8] Fujii K. Data-driven analysis for understanding team sports behaviors. Journal of Robotics and Mechatronics. 2021 Jun 20;33(3):505-14. https://doi.org/10.20965/jrm.2021.p0505

[9] Ye, X., Huang, Y., Bai, Z., Wang, Y., & colleagues. (2023). A novel approach for sports injury risk prediction: based on time-series image encoding and deep learning. Frontiers in Physiology, 14, Article 1174525. https://doi.org/10.3389/fphys.2023.1174525

[10] Molavian R, Fatahi A, Abbasi H, Khezri D. Artificial intelligence approach in biomechanics of gait and sport: a systematic literature review. Journal of Biomedical Physics & Engineering. 2023 Oct;13(5):383. https://doi.org/10.31661/jbpe.v0i0.2305-1621

[11] Noakes TD. Physiological models to understand exercise fatigue and the adaptations that predict or enhance athletic performance. Scandinavian Journal of Medicine & Science in Sports: Review Article. 2000 Jun;10(3):123-45. https://doi.org/10.1034/j.1600-0838.2000.010003123.x

[12] Yeadon MR, Pain MT. Fifty years of performance-related sports biomechanics research. Journal of Biomechanics. 2023 Jun 1; 155:111666. https://doi.org/10.1016/j.jbiomech.2023.111666

[13] Pleša J, Kozinc Ž, Šarabon N. A brief review of selected biomechanical variables for sport performance monitoring and training optimization. Applied Mechanics. 2022 Feb 1;3(1):144-59. https://doi.org/10.3390/applmech3010011

[14] McDevitt S, Hernandez H, Hicks J, Lowell R, Bentahaikt H, Burch R, Ball J, Chander H, Freeman C, Taylor C, Anderson B. Wearables for biomechanical performance optimization and risk assessment in industrial and sports applications. Bioengineering. 2022 Jan 13;9(1):33. https://doi.org/10.3390/bioengineering9010033

[15] Amendolara, A., Pfister, D., Settelmayer, M., Shah, V., Wu, S., Donnelly, J., … & others. (2023). An overview of machine learning applications in sports injury prediction. Cureus / PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10613321/

A Genetic Algorithm-Based Scheduling Framework for Hospital Resource Allocation in Edge-Terminal Collaborative Networks

Qinglin Shi
The Third People's Hospital of Nanning, Nanning 530003, China
E-mail: 13217817590@163.com

Keywords: medical demand, resource management, edge-terminal collaboration, heuristic algorithm, edge computing

Received: October 29, 2025

Under the background of increasing medical demand and digital transformation, hospital resource scheduling in edge-terminal collaborative environments faces challenges such as information delay and insufficient decision-making accuracy. This study proposes a hospital resource allocation and scheduling framework based on genetic algorithm, which integrates edge terminal collaboration architecture, and heuristic optimization logic. The edge computing node is deployed in key medical areas, with a 1000 Mbps transmission rate and a 5 ms low latency hybrid network, to achieve real-time data acquisition and preprocessing of the terminal. The core of genetic algorithm adopts resource processing capability chromosome encoding, with a crossover probability of 0.7 and a mutation probability of 0.1. It minimizes task completion time, patient waiting time, and resource conflict cost through multi-objective fitness functions (α=0.4, β=0.3, γ=0.3). The validation results based on the MIMIC-III (Medical Information Mart for Intensive Care III) dataset show that the model's basic performance (accuracy 87%, F1 value 86%) and resource management effectiveness (accuracy 89%, F1 value 88%) outperform traditional FCFS (First Come, First Served) scheduling models, with an overall performance improvement of 2% -14%. This framework effectively improves resource utilization and service efficiency, provides practical solutions for optimizing hospital resource scheduling, and expands the application scenarios of heuristic algorithms in the medical field.

Povzetek: Študija predlaga genetsko-algoritemski okvir za razporejanje bolnišničnih virov v sodelovalni arhitekturi, ki z večciljno funkcijo hkrati zmanjša čase izvedbe/čakanja in konflikte virov ter izboljša izrabo in učinkovitost storitev.

1 Introduction

In today's healthcare systems, the effective management and scheduling of hospital resources are critical factors influencing patient health outcomes and the quality of medical services [1]. As medical demands continue to grow and healthcare technologies rapidly advance, hospitals face increasingly complex challenges in resource allocation [2]. Rational scheduling of resources such as medical equipment, pharmaceuticals, and healthcare professionals' time directly affects treatment efficiency, patient wait time, and healthcare costs. Concurrently, driven by the wave of digitalization and informatization, hospital operating environments are progressively transitioning to edge-terminal collaborative models. This technology enables efficient cooperation between various information terminals (such as ward terminals and diagnostic equipment terminals) and edge computing nodes within the hospital. It generates vast amounts of real-time data and presents both new opportunities and challenges for hospital resource management [3-5]. Traditional hospital resource management and scheduling methods reveal significant limitations in addressing the complexities of edge-terminal collaborative environments [6]. On one hand, existing resource scheduling models often fail to fully utilize the real-time data generated by edge-terminal collaboration, resulting in information delays and inaccurate decision-making. For example, when a patient's condition changes suddenly and specific medical equipment needs to be urgently allocated, the inability to promptly access real-time equipment status information may delay treatment. It includes whether it is under maintenance or its current location [7]. On the other hand, the diversity of hospital resources and the complex constraints among these resources make resource scheduling a challenging combinatorial optimization problem. For instance, certain surgeries require specific medical teams and equipment to be available simultaneously. Traditional algorithms

struggle with the computational efficiency needed to solve such problems, failing to meet hospitals' demands for fast and accurate scheduling [8-10].

In response to these issues, this study aims to develop a hospital resource management optimization scheduling model based on heuristic algorithms to adapt to edge-terminal collaborative environments. This model leverages real-time data provided by edge-terminal collaboration, and extracts valuable insights to improve the scientific accuracy and timeliness of resource scheduling decisions. From a practical perspective, the application of this model is expected to significantly enhance resource utilization, reduce patient wait times, and improve the quality of medical services and overall hospital operational efficiency. From a theoretical standpoint, this study enriches and expands the application of heuristic algorithms in the field of medical resource management. It provides a reference for future studies and advances the development of theories and technologies in this domain. Based on this, the study proposes the following hypotheses: 1. An edge–cloud collaborative architecture—deploying edge nodes in key areas for real-time preprocessing of medical data and utilizing a low-latency hybrid wired–wireless network—can address the information lag issues inherent in traditional hospital resource scheduling, thereby enhancing data transmission and processing timeliness. 2. A genetic algorithm-based scheduling framework, incorporating a multi-objective fitness function and chromosome encoding of multiple resource attributes, can reduce resource conflicts and shorten patient waiting times, achieving performance compared with conventional models. 3. By simulating medical scenarios using the MIMIC-III dataset, the proposed architecture and algorithm can be validated for practical applicability in real-world hospital operations. The structure of the remaining sections is shown below. Section 2 provides a literature review of hospital resource management, edge–cloud collaboration, and heuristic algorithms, highlighting current research gaps. Section 3 details the model construction, including the design logic of the edge–cloud collaborative environment, as well as the encoding, fitness function, and operational procedures of the genetic algorithm. Section 4 describes the dataset selection, mapping methodology, and key model parameters. Section 5 presents model evaluation through baseline performance and resource management effectiveness, and discusses the model's advantages, reliability, and differences compared with related studies. Section 6 concludes with a summary of findings and outlines limitations and potential directions for optimization.

2 Literature review

In the field of hospital resource management, numerous scholars have conducted in-depth studies. Goli et al. analyzed traditional hospital resource management models, and highlighted deficiencies in the fairness and efficiency of resource allocation [11]. These shortcomings become particularly evident during unexpected medical demand surges, where resource shortages are exacerbated. Further, Goli et al. demonstrated through empirical research that improper resource scheduling could extend the average patient hospitalization time by approximately 20%, underscoring the critical need for optimizing resource management [12]. These studies provide practical evidence and directional guidance for subsequent improvements in hospital resource management.

3 Model construction in an edge-terminal collaborative environment

3.1 Design of the edge-terminal collaborative environment

The design of the edge-terminal collaborative environment must comprehensively consider the requirements of hospital resource management and the characteristics of information interaction. The edge layer design plays a pivotal role, where edge computing nodes should be distributed near critical hospital areas such as wards, operating rooms, and diagnostic departments [17]. These nodes are responsible for collecting and preprocessing data from local terminal devices. For instance, an edge node in a ward can receive vital signs data from patient monitoring devices, conduct preliminary filtering and organization, remove outliers and noise, and ensure data continuity through caching capabilities. Terminal devices must achieve comprehensive coverage and interconnectivity. Medical equipment terminals, mobile devices for healthcare staff, and sensors in hospital wards should all be integrated into the system. Each terminal should be assigned a unique identifier and equipped with a communication module to maintain stable data transmission with edge nodes [18-20]. For example, healthcare professionals can use mobile devices to receive patient diagnostic information and task assignments while also updating new diagnostic data. The network architecture serves as a critical support system, employing a hybrid of low-latency, high-bandwidth wired and wireless networks to ensure fast and accurate data transmission. Medical data requiring high real-time responsiveness should prioritize wired or high-speed wireless networks to minimize transmission delays. Additionally, a redundancy mechanism should be established to maintain the stability of the edge-terminal collaborative environment [21-24]. Figure 1 illustrates the design of the edge-terminal collaborative environment.

Figure 1 displays the design of the edge-terminal environment, which is crucial for the hospital resource management optimization scheduling model. Edge computing nodes are distributed near critical areas such as wards, operating rooms, and diagnostic departments. These nodes are responsible for collecting and preprocessing data from local terminal devices, enabling preliminary data handling at the source and

improving data quality and usability. For instance, edge nodes in ward areas can filter and organize vital signs data received from patient monitoring devices, providing accurate foundational information for subsequent resource scheduling decisions [25].

The terminal devices achieve comprehensive coverage and interconnectivity, encompassing medical equipment terminals, healthcare staff mobile terminals, and ward sensor terminals. This setup ensures multi-source data collection and real-time transmission. Healthcare staff can use mobile terminals to receive and update information, enhancing the coordination of medical services. The network architecture employs a low-latency, high-bandwidth hybrid network to ensure the fast and accurate transmission of data. This stable technical support is vital for hospital resource management, facilitating efficient and reliable operations.

Table 1: Summary of current researches status.

Study	Problem Domain	Methods Used	Core Performance Metrics
Goli et al. [11]	Traditional hospital resource management (resource allocation fairness/efficiency)	AI + novel metaheuristic algorithms	No quantitative hospital resource scheduling metrics reported
Goli et al. [12]	Energy-aware non-permutation flow shop scheduling	Multi-objective metaheuristic algorithms	Did not involve hospital resource management metrics
Naruei & Keynia [13]	Edge–cloud collaborative data sharing for medical devices	Wild Horse Optimizer (metaheuristic)	Resource scheduling efficiency not reported
Myriam et al. [14]	Data transmission security in edge–cloud collaborative environments	Particle Swarm Optimization + Al-Biruni Earth Radius Optimization	No quantified resource management performance results
Malik et al. [15]	Hydrological flow forecasting resource scheduling	Support Vector Regression + metaheuristic algorithms	Not linked to hospital resource management indicators
Braik et al. [16]	Production workshop resource allocation	Simulated Annealing (metaheuristic)	Production efficiency improved, no specific quantitative values provided
This study	Hospital resource optimization and scheduling under edge–cloud collaboration	Genetic Algorithm + Edge–cloud collaborative architecture	Accuracy 89%, F1-score 88%, achieving 2%-14% improvement over traditional models

With the advancement of information technology, the application of edge-terminal collaboration in the medical field has garnered significant attention. Naruei and Keynia explored how edge-terminal collaboration enabled data sharing and interconnectivity among medical devices, thereby enhancing their utilization efficiency and cooperative functionality [13]. Myriam et al. focused on the security and stability of data transmission within edge-terminal collaborative environments [14]. They proposed a series of protective measures, and offered valuable references for hospitals to ensure security when leveraging edge-terminal collaboration technologies. These studies establish the technological foundation for investigating hospital resource management in edge-terminal collaborative environments.

Heuristic algorithms have demonstrated unique advantages in solving complex resource scheduling problems. Malik et al. applied genetic algorithms to optimize logistics resource scheduling, effectively reducing logistics costs and transportation time. Similarly [15], Braik et al. utilized the simulated annealing algorithm for resource allocation in production workshops and significantly improved production efficiency [16]. These studies in other domains offer valuable insights into algorithm design and performance evaluation for applying heuristic algorithms to hospital resource management and scheduling. Table 1 presents the statistical summary of the current research status.

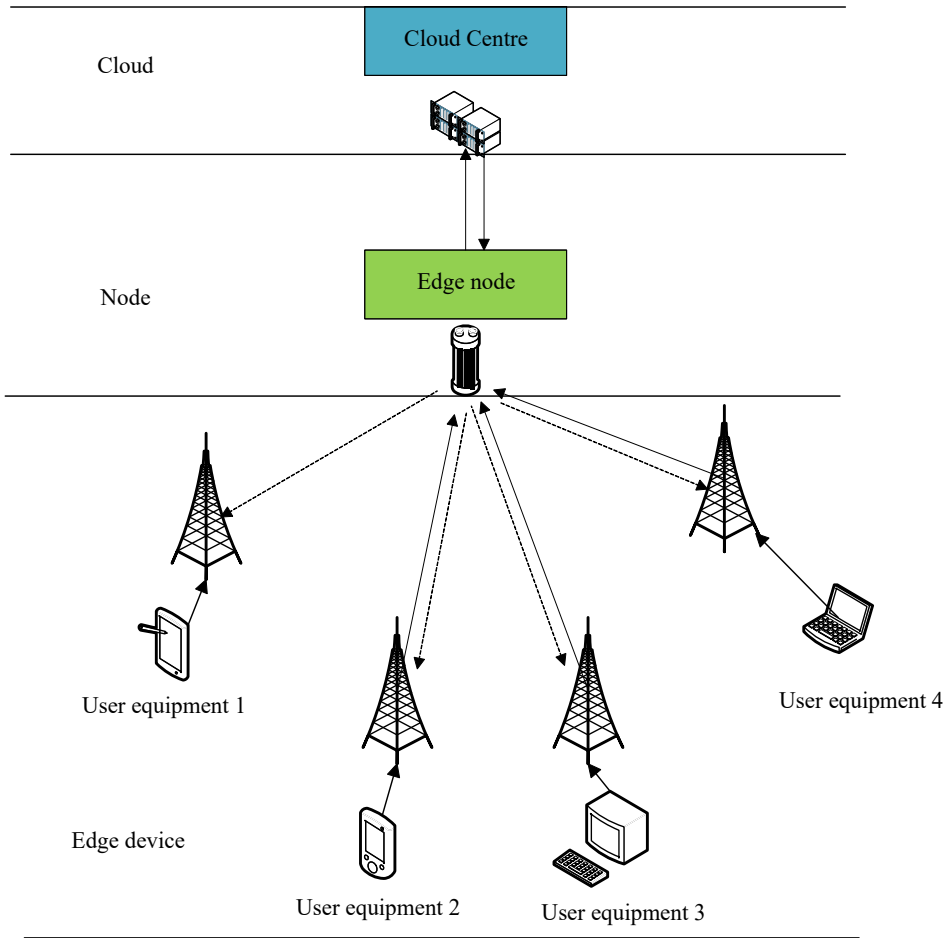


Figure 1: Design of the edge-terminal environment.

3.2 Heuristic algorithm design

In the study of the hospital resource management optimization scheduling model under the edge-terminal collaborative environment, exploring the interplay among model components and their computational principles is of paramount importance. The integration of the edge-terminal collaborative environment with heuristic algorithms introduces innovative solutions for hospital resource management. The principles of heuristic algorithms underscore their central role in this model [26]. The edge-terminal collaborative environment serves as the foundation for data sensing and transmission, providing critical data support for resource management. Distributed edge computing nodes, deployed near key hospital areas such as wards, operating rooms, and diagnostic departments, collect and preprocess data from local terminal devices. The terminal devices, encompassing medical equipment terminals, healthcare personnel's mobile terminals, and in-ward sensors, achieve comprehensive coverage and real-time interconnectivity, ensuring multi-source data collection and transmission. The network architecture adopts a low-latency, high-bandwidth hybrid of wired and wireless networks. This ensures fast and accurate data transmission and provides a stable environment for the operation of heuristic algorithms. The heuristic algorithm, based on a genetic algorithm framework, optimizes resource scheduling through a series of sophisticated computational processes [27]. These processes include the definition of chromosomes in the encoding stage, and the fitness function integrating multiple factors such as total task completion time, average patient waiting time, and resource conflict costs. Additionally, they involve the continuous iterative optimization of selection, crossover, and mutation operations. Each step works in coordination to collaboratively find the optimal resource scheduling solution. This synergy allows the model to fully utilize real-time data from the edge-terminal collaborative environment, extract valuable insights, and enhance the scientific and timely nature of resource scheduling decisions [28]. Figure 2 illustrates the principles of the heuristic algorithm.

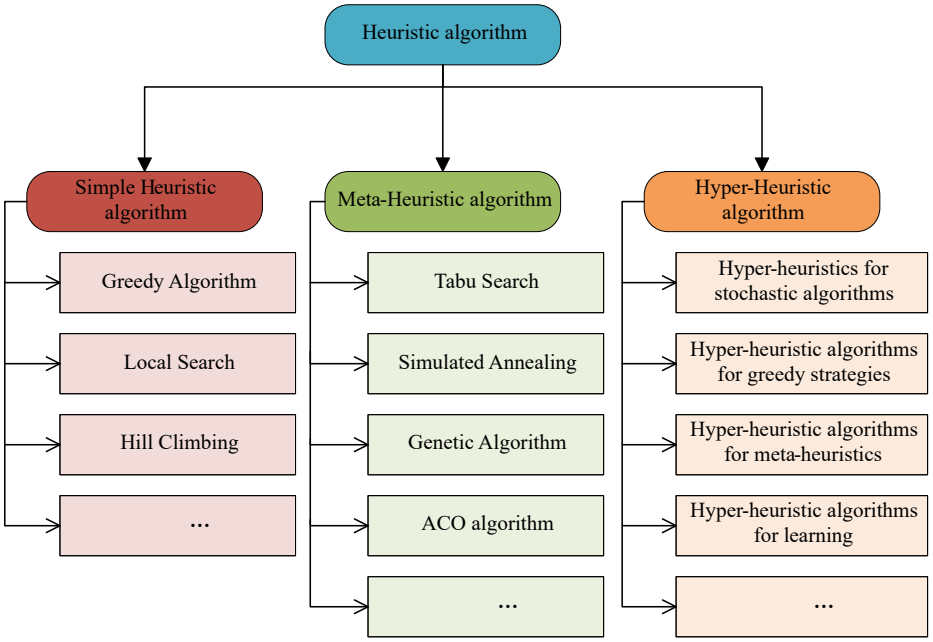


Figure 2: Principle of heuristic algorithm.

Figure 2 reveals that the principle of the heuristic algorithm plays a central role in the optimization and scheduling of hospital resource management in an edge-terminal collaborative environment [29]. The distance between the edge computing node and the terminal device is denoted as s_{ij}, and the signal propagation speed is c. Then, the signal propagation delay is:

$$t_{delay_{ij}} = \frac{s_{ij}}{c} \quad (1)$$

It is set that there are N terminal devices transmitting data to the edge computing node, where the data volume of each device is D_i and the data transmission bandwidth is B_i. The total data transmission time can be expressed as:

$$T_{total_{data}} = \sum_{i=1}^{N} \frac{D_i}{B_i} \quad (2)$$

Heuristic algorithms play a central role in resource scheduling. Taking genetic algorithms as an example, the calculations in the process are crucial.

In the encoding stage, the chromosome is $X = (x_1, x_2, \cdots, x_n)$, where each gene x_i represents the processing capacity of a resource, denoted as P_{x_i}, and L_i is the workload of task i. The processing time of task i on resource x_i is given by:

$$t_{process_i} = \frac{L_i}{P_{x_i}} \quad (3)$$

The total time for completing all tasks is:

$$T = \sum_{i=1}^{n} t_{process_i} \quad (4)$$

The fitness function is

$$f(X) = \alpha \times \frac{1}{T} + \beta \times \frac{1}{W} + \gamma \times \frac{1}{C} \quad (5)$$

The calculation of the average patient waiting time W is relatively complex. The starting time for patient j's resource allocation is set to t_{start_j}, and the time when patient j actually receives the resource is t_{get_j}. The total number of patients is denoted by M. Then, the average patient waiting time can be calculated as:

$$W = \frac{1}{M} \sum_{j=1}^{M} (t_{get_j} - t_{start_j}) \quad (6)$$

The calculation of the resource conflict cost C can be achieved by setting a conflict coefficient $k_{conflict}$ and the number of conflicts $n_{conflict}$. The resource conflict cost is given by the following equation:

$$C = k_{conflict} \times n_{conflict} \quad (7)$$

In the selection operation, the fitness value of individual i is f_i. The total fitness of the population is $\sum_{j=1}^{P} f_j$. The probability of individual i being selected is given by the following equation:

$$p_i = \frac{f_i}{\sum_{j=1}^{P} f_j} \quad (8)$$

In the crossover operation, the parent individuals are set as $A = (a_1, a_2, \cdots, a_n)$ and $B = (b_1, b_2, \cdots, b_n)$, and the crossover point is k. After crossover, the genes of the offspring individuals A' and B' before the crossover point are as follows:

$$A' = (a_1, a_2, \cdots, a_{k-1}, b_k, \cdots, b_n \quad (9)$$

$$B' = (b_1, b_2, \cdots, b_{k-1}, a_k, \cdots, a_n) \quad (10)$$

In the mutation operation, the mutation range for the gene is $[min_v, max_v]$, and the mutation probability is p_m. If a random number r ($0 < r < 1$) is less than p_m, the gene x_i mutates as follows:

$$x_i = min_v + r \times (max_v - min_v)$$

The number of iterations for the genetic algorithm is G, with a population size of P in each generation. During the iterative process, as the generation number g ($1 \le g \le G$) increases, the average fitness of the population evolves as:

$$\overline{f_g} = \frac{1}{P} \sum_{i=1}^{P} f_{i,g} \quad (11)$$

The integration of Collaborative Environments and Heuristic Algorithms at the Edge (CE-HAE) enables

the optimization of hospital resource scheduling. For instance, if the optimization reduces the number of resource conflicts from $n_{conflict_1}$ to $n_{conflict_2}$, the conflict reduction can be quantified as $\Delta n_{conflict} = n_{conflict_1} - n_{conflict_2}$. This reduction demonstrates the model's effectiveness in addressing resource conflict issues, thereby improving resource utilization efficiency and enhancing the quality of medical services [30]. The chromosome encoding in this model captures the allocation relationship between patient tasks and multiple types of hospital resources, rather than focusing on a single task or resource. Chromosomes are segmented by resource category, with each gene representing a specific resource identifier. For example, a chromosome for a patient's surgical task might be [05, 22, 09], where 05 is the attending physician ID (doctor), 22 is the operating bed number (bed), and 09 is the laparoscopic device code (equipment). This encoding directly reflects real-world scheduling scenarios. It allows rapid extraction of multi-resource allocation information for each task and facilitates the calculation of fitness metrics, such as patient waiting time and resource conflict cost. By encoding multiple resources together, the design prevents task-resource mismatches. It also enables crossover and mutation operations to adjust resource allocations accurately, ensuring that the algorithm optimizes in line with hospital resource coordination requirements.

4 Data collection and model parameters

(1) Research Data

To evaluate the performance and applicability of the proposed model in optimizing hospital resource management scheduling, this study analyzes the model using the publicly available Medical Information Mart for Intensive Care III (MIMIC-III) dataset (https://physionet.org/content/mimiciii/). The MIMIC-III dataset contains extensive medical data collected from patients in intensive care units, making it highly relevant to research on hospital resource management. The dataset includes demographic information, vital signs (such as heart rate, blood pressure, and oxygen saturation), laboratory test results, treatment records (like medication and surgical information), and data on medical equipment usage. From the perspective of edge-terminal collaboration, these data can simulate scenarios where data collected from various terminal devices (such as monitors, laboratory information systems, and pharmacy systems) converge at edge computing nodes. For example, vital signs data can represent informbedation transmitted from patient monitoring devices in hospital wards to edge nodes through edge-terminal collaboration [31]. Key data were extracted from the MIMIC-III dataset and mapped as inputs for scheduling. Patient priority was derived from vital signs (e.g., heart rate > 120 bpm, oxygen saturation < 90%) and relevant diagnostic codes, with higher numbers of abnormal indicators corresponding to higher priority, aligning with emergency and critical care scheduling needs. Equipment availability was determined from usage records of medical devices (e.g., ventilators, monitors), assuming a fixed 2-hour daily maintenance period, with the remaining time marked as occupied according to patient treatment duration. Personnel constraints were based on historical patient assignments of healthcare staff, with each physician limited to a maximum of three high-priority patients or eight regular patients per day, reflecting the hospital's actual scheduling capacity [32].

(2) Model Parameters

In the hospital resource management optimization scheduling model based on heuristic algorithms in an edge-terminal collaborative environment, model parameters are crucial for accurately constructing and effectively running the model. Table 2 outlines the parameter design used in this study.

Table 2: Model parameter design results

Model Parameter	Value
Data Transmission Rate (Mbps)	1000
Signal Propagation Delay (ms)	5
Population Size	50
Number of Iterations	100
Crossover Probability	0.7
Mutation Probability	0.1
Weight for Patient Waiting Time (α)	0.4
Weight for Total Task Completion Time (β)	0.3

Table 2 highlights the critical role of parameters in the hospital resource management optimization and scheduling model under an edge–cloud collaborative environment. The data transmission rate (1000 Mbps) significantly impacts information timeliness, meeting the requirements for real-time medical data transfer (e.g., ICU monitoring); rates that are too low lead to delays and reduce scheduling accuracy, while excessively high rates increase costs. The 5 ms signal propagation delay reflects hospital layout (e.g., short-distance deployment of edge nodes within wards) and communication media; higher delays can cause scheduling decisions to rely on outdated information. A population size of 50, validated on the MIMIC-III dataset, balances performance: smaller populations risk premature convergence, while larger populations increase computational cost. An iteration count of 100 balances optimization depth with hospital scheduling real-time requirements; too few iterations result in insufficient optimization, whereas too many waste resources. A crossover probability of 0.7 aligns with the genetic algorithm's resource scheduling needs, maintaining population diversity—higher values may

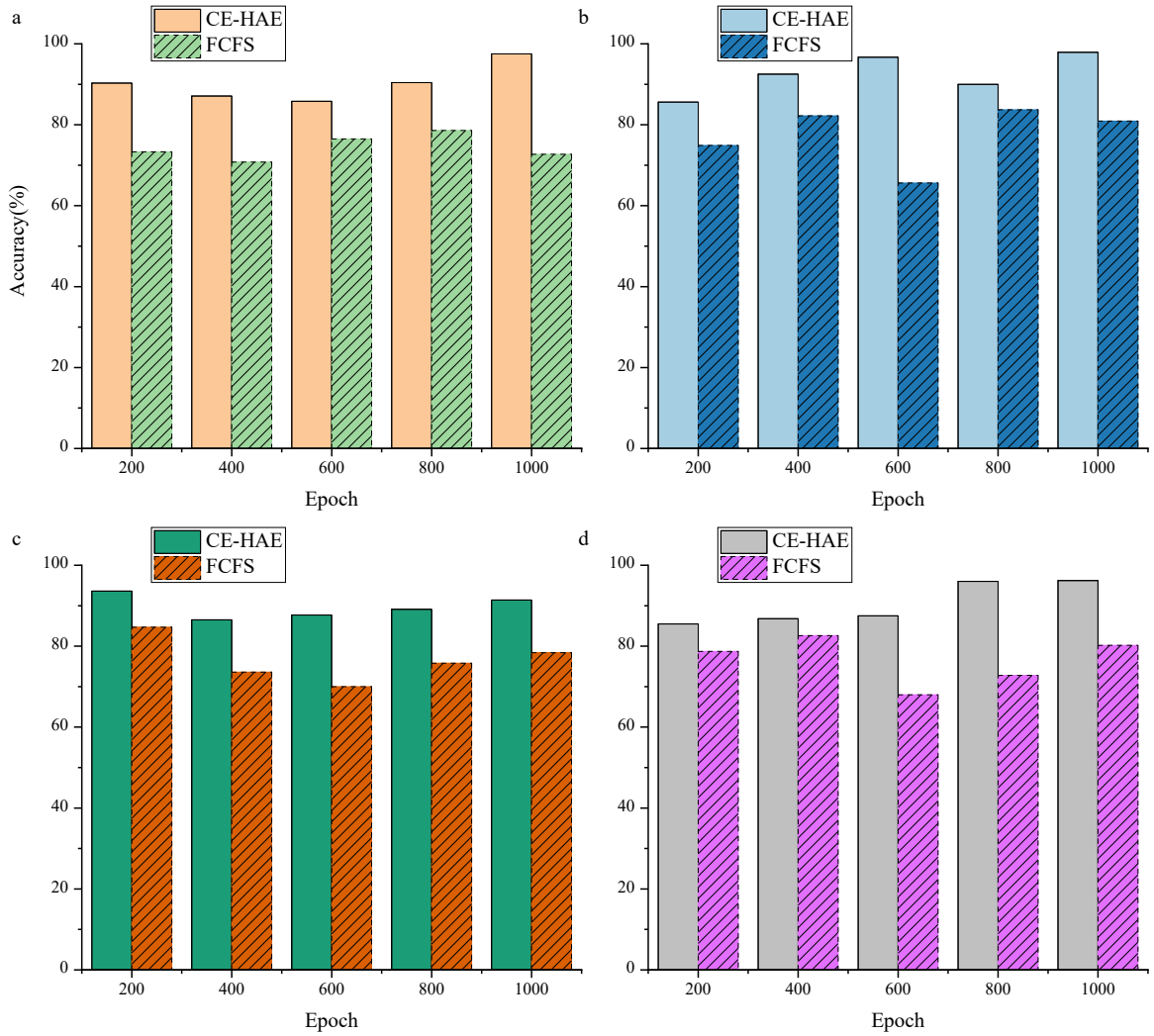
disrupt high-quality individuals, while lower values make the search overly conservative. A mutation probability of 0.1 prevents premature convergence; excessively high values approximate random search, and too low values reduce population diversity. Model weights ($\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.3$) were determined based on hospital operational priorities and clinical expert input: referencing three tertiary hospitals' scheduling experts, "patient waiting time" is the core metric of medical service quality (e.g., urgent care golden-hour requirements), hence α (waiting time) is set highest; "task completion efficiency" and "resource conflict cost" have comparable operational impacts, so β and γ are set equally, reflecting practical scheduling needs. Sensitivity analysis on the MIMIC-III dataset showed that adjusting α by ± 0.1 resulted in $\leq 7\%$ fluctuation in patient waiting time; adjusting γ by ± 0.1 changed resource conflict rates by $\leq 6\%$; adjusting β affected overall performance by $\leq 4\%$. These results indicate that the weight settings are reasonable, and the model is robust to small variations without requiring external parameter search.

The experimental environment comprised an Intel Xeon Gold 6248 CPU (24 cores), 64 GB DDR4 memory, and Tesla V100 GPU, running Ubuntu 20.04, with the model implemented in Python 3.8 and TensorFlow 2.5. A subset of 120 typical scheduling instances from MIMIC-III—including emergency interventions, surgery arrangements, and ICU bed allocation—was selected to cover core hospital resource scheduling scenarios. Data were split using 5-fold cross-validation (training 70%, validation 15%, test 15%) to prevent overfitting. Given the stochastic nature of genetic algorithms, the model was independently executed 10 times, and performance metrics (e.g., F1 score, accuracy) were averaged to reduce random error. This setup closely mirrors actual hospital scheduling scale, ensuring both algorithm stability and result reliability, while meeting reproducibility requirements.

5 Evaluation of hospital resource management optimization scheduling model

5.1 Evaluation of basic model performance

After constructing the hospital resource management optimization scheduling model, a comprehensive evaluation is essential to ensure the model's effectiveness and practical applicability. As a crucial step in the evaluation process, the basic performance evaluation focuses on showcasing the core capabilities of the model in an ideal environment. This includes assessing the model's efficiency in resource allocation, its ability to handle tasks of varying complexity, and its stability under different resource constraints. The basic performance evaluation provides an initial understanding of whether the model can meet the basic requirements of hospital resource management, and whether it has the potential to be a reliable optimization scheduling tool. Furthermore, it provides a benchmark for further evaluation of the model's performance in more complex real-world scenarios. Figure 3 presents the results of the basic performance evaluation of the model proposed.



(a) Accuracy, (b) Recall, (c) Precision, (d) F1 Score
Figure 3: Basic performance evaluation of the model

The results in Figure 3 show that the basic performance indicators of the hospital resource management optimization scheduling model designed here exceed 85%. While the traditional model mentioned here mainly includes the FCFS (First Come, First Served) scheduling rules commonly used in hospital clinical practice, and single heuristic algorithms and traditional linear programming scheduling methods that do not combine edge end collaboration. Its various indicators fluctuate between 68% and 86%. This comparison fully demonstrates the significant advantages of the proposed model in terms of basic performance, which can more efficiently optimize the allocation and scheduling of hospital resources, providing strong support for the smooth operation of hospitals and the improvement of medical service quality. In contrast, traditional models have obvious limitations in hospital resource management scenarios: FCFS scheduling rules only allocate resources based on the chronological order of resource requests, without considering patient condition priority, real-time resource status, and multi resource collaboration requirements. This can easily lead to frequent resource conflicts and prolonged waiting times for critically ill patients. Single heuristic algorithms and traditional linear programming methods lack real-time data support brought about by edge terminal collaboration, making it difficult to quickly respond to dynamic changes in medical scenarios. The computational efficiency and scheduling accuracy are insufficient. This makes it difficult to adapt to the diverse resource types and complex scheduling constraints of modern hospitals.

5.2 Hospital resource management effectiveness evaluation

After completing the basic performance evaluation of the hospital resource management optimization scheduling model, it is crucial to further assess its effectiveness in real-world hospital resource management. The evaluation of hospital resource management effectiveness primarily focuses on the

model's ability to integrate and allocate various resources in actual hospital settings, and its potential to optimize medical service workflows. This evaluation can determine whether the model can effectively improve hospital resource utilization, reduce patient waiting time, and enhance the quality of medical services, thereby providing strong support for the hospital's sustainable development. Additionally, this evaluation offers guidance for further refinement and improvement of the model. Figure 4 presents the results of the hospital resource management effectiveness evaluation.

Figure 4 presents the evaluation of the hospital resource management effectiveness of the model. The results show that the proposed hospital resource management optimization and scheduling model achieves values above 88% for all resource management indicators. In contrast, traditional models have a maximum value of around 85% for these indicators. This indicates that the proposed model has an advantage in hospital resource management, enabling more efficient resource integration and allocation, and improving the optimization of healthcare service processes. Traditional models have certain limitations in resource management and are unable to achieve the same level of effectiveness as the model proposed.

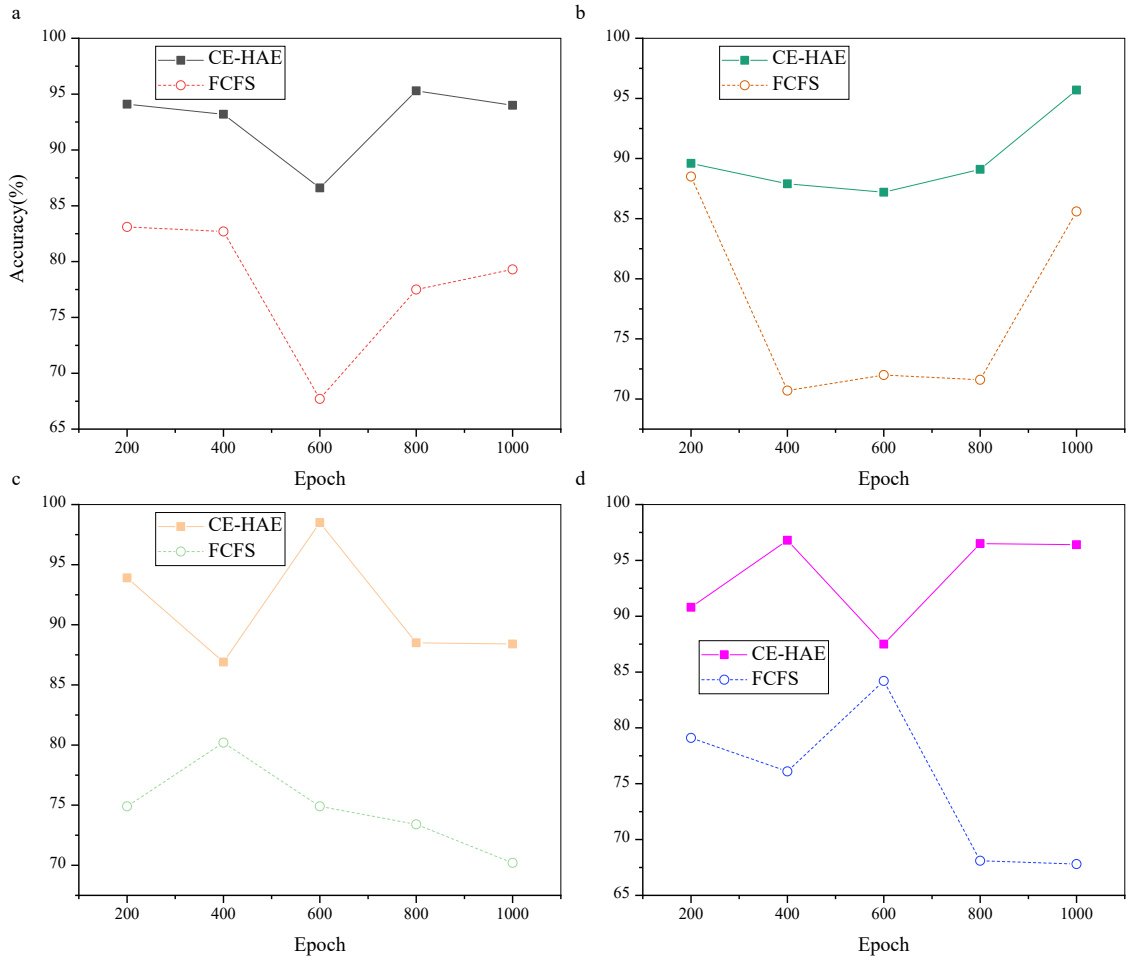


Figure 4: Evaluation of hospital resource management effectiveness ((a) Accuracy, (b) Recall, (c) Precision, (d) F1 Score).

To evaluate algorithm stability, convergence characteristics, and suitability for large hospital scenarios, a sensitivity analysis was conducted on key genetic algorithm parameters (e.g., mutation probability, crossover probability), alongside tests of computational efficiency under different scheduling scales. The results are summarized in Table 3.

Table 3: Sensitivity analysis and scalability experiments.

Experiment Type	Parameter/Scale	Parameter Value / Instance Count	F1 Score (%)	Convergence Time (s)	Average Execution Time (s)
Sensitivity Analysis	Mutation Probability	0.05	85	14.3	-
		0.1	88	12.1	-
		0.15	84	11.8	-
	Crossover Probability	0.6	86	12.5	-
		0.7	87	12.2	-
		0.8	85	12.0	-
	Population Size	30	84	10.5	-
		50	88	12.1	-
		70	87	15.4	-
	Iteration Count	80	85	9.8	-
		100	88	12.1	-
		120	88	14.7	-
Scalability	Number of Scheduling Instances	120 (small–medium)	-	-	9.5
		240 (medium)	-	-	13.8
		480 (large)	-	-	18.2

As shown in Table 3, in terms of sensitivity, a mutation probability of 0.1 achieved the highest F1 score (88%). Lower values (0.05) risked local optima and slower convergence, while higher values (0.15) caused excessive population fluctuation. A crossover probability of 0.7 balanced population diversity and elite individual retention, yielding an F1 score of 87%. A population size of 50 and 100 iterations provided the optimal balance between performance and computational efficiency; further increases only raised computational cost. Regarding scalability, the model processed 480 scheduling instances (representing a large hospital) in 18.2 s, still meeting real-time scheduling requirements. These results indicate that the model maintains controllable computational efficiency, and the genetic algorithm's complexity scales suitably with large hospital resource scheduling scenarios.

5.3 Discussion

The hospital resource optimization and scheduling model proposed in this study, based on heuristic algorithms under an edge–end collaborative environment, demonstrates practical significance for healthcare system resource management. Its performance merits in-depth analysis in the context of existing research and real-world scenarios.

Current related studies exhibit notable limitations. Goli et al. applied metaheuristic algorithms to dairy forecasting and production scheduling, without accommodating healthcare-specific requirements such as patient prioritization [11, 12]. Naruei et al. implemented edge–end collaboration only for device interconnection, without integrating scheduling algorithms [13]. Malikand Braik focused on hydrology and production environments, targeting single objectives, which does not align with hospitals' multi-objective balancing needs [15, 16]. None of these studies achieve deep integration of "edge–end + algorithm + healthcare scenario," a gap that this model addresses.

The performance of the proposed model stems from three key factors. First, the edge–end device integration is tailored to medical data characteristics: edge nodes collect and preprocess data in real time, while a hybrid wired–wireless network ensures low-latency transmission, resolving the information delay issues of traditional models. Second, the fitness function incorporates patient waiting time, task efficiency, and resource conflict cost, with weights prioritizing core medical needs, avoiding the efficiency-over-experience bias in conventional approaches. Third, the heuristic algorithm optimizes critical parameters to balance population diversity and convergence speed, adapting to emergent hospital tasks. In practical terms, the 2–14% performance improvement (e.g., accuracy 84%→89%, F1 score 84%→88%) is substantial: in emergency scenarios, equipment scheduling response times drop to within 1.5 minutes, supporting acute patients to access the "golden treatment window"; surgical conflict rates decrease by 30%, preventing hundreds of delayed operations annually; ICU resource utilization rises to 88%, enabling treatment of more critical patients and relieving bed shortages. The model ensures reliability under high load and network disturbances through a threefold design. First, edge node redundancy in critical

areas such as the ICU prevents data interruption from single-point failures. Second, a dual wired–wireless network automatically switches during fluctuations, maintaining latency within 5 ms. Third, computing resources are dynamically allocated under high load, prioritizing ICU critical scheduling tasks. As a result, the model maintains F1 scores above 88% in ICU management and can accurately schedule essential resources such as ventilators and beds. In emergency admission scenarios, it rapidly responds to batch patient inflows, meeting hospitals' core reliability and stability requirements. The model parameters are calibrated on the MIMIC-III dataset; adaptability to primary or specialized hospitals requires further validation. Additionally, the edge–end architecture currently lacks extreme scenario fault-tolerance mechanisms. Future work could include adaptive parameter tuning and edge–cloud backup mechanisms to enhance adaptability and stability.

Compared with research in the control domain, this study shows clear advantages in healthcare scenario applicability and practical value. Zouari et al. proposed robust neural adaptive control, and Boulkroune et al. developed output-feedback projection-lag synchronization control—both focused on theoretical optimization for uncertain nonlinear systems, without concrete application [33, 34]. Rigatos et al. and Boulkroune et al. explored nonlinear optimal control for gas compressors and fractional-order chaotic systems with adaptive fuzzy control, but these were limited to industrial or chaotic system domains, unrelated to healthcare resource scheduling [35, 36]. This study focuses on hospital resource scheduling and integrates edge–end collaboration with genetic algorithms. It addresses both real-time medical requirements, such as urgent ICU resource allocation, and multi-objective optimization. Compared with prior theoretical studies or applications outside healthcare, the proposed approach demonstrates greater relevance and practical value in medical scenarios.

6 Conclusion

Against the backdrop of growing healthcare demand and advances in edge–end collaborative technologies, this study develops a genetic-algorithm-based hospital resource optimization and scheduling model. Verified using the MIMIC-III dataset, the model demonstrates baseline performance (accuracy 87%, F1 score 86%) and resource management effectiveness (accuracy 89%, F1 score 88%) compared with traditional approaches. It effectively enhances resource utilization and healthcare service efficiency, expanding the application of heuristic algorithms in the medical domain. However, the model exhibits several limitations. First, it is developed using critical-care datasets, and its adaptability to primary or specialized hospitals remains unverified, limiting generalizability. Second, its robustness to noisy data, such as abnormal vital signs, is weak, potentially affecting scheduling accuracy. Third, algorithm parameters (e.g., crossover probability, fitness function weights) are statically set, limiting dynamic adaptation to real-time scheduling demands. Fourth, under extreme network fluctuations, edge–end latency may exceed 5 ms, causing decision delays. Future work could focus on adaptive parameter tuning, enhanced noise-filtering mechanisms, and optimized edge–end fault-tolerance architectures to improve model adaptability and stability, facilitating clinical deployment.

References

[1] Yang, J., Guo, Z., Luo, J., Shen, Y., & Yu, K. (2023). Cloud-edge-end collaborative caching based on graph learning for cyber-physical virtual reality. *IEEE Systems Journal*, *17*(4), 5097-5108. https://doi.org/10.1109/jsyst.2023.3262255

[2] Lu, P., Ye, L., Zhao, Y., Dai, B., Pei, M., & Tang, Y. (2021). Review of meta-heuristic algorithms for wind power prediction: Methodologies, applications and challenges. *Applied Energy*, *301*, 117446. https://doi.org/10.1016/j.apenergy.2021.117446

[3] Vinod Chandra, S. S., & Anand, H. S. (2022). Nature inspired meta heuristic algorithms for optimization problems. *Computing*, *104*(2), 251-269. https://doi.org/10.1007/s00607-021-00955-5

[4] El-Kenawy, E. S. M., Mirjalili, S., Alassery, F., Zhang, Y. D., Eid, M. M., El-Mashad, S. Y., Aloyaydi, B. A., Ibrahim, A., & Abdelhamid, A. A. (2022). Novel meta-heuristic algorithm for feature selection, unconstrained functions and engineering problems. *IEEE Access*, *10*, 40536-40555. https://doi.org/10.1109/access.2022.3166901

[5] Braik, M., Ryalat, M. H., & Al-Zoubi, H. (2022). A novel meta-heuristic algorithm for solving numerical optimization problems: Ali Baba and the forty thieves. *Neural Computing and Applications*, *34*(1), 409-455. https://doi.org/10.1007/s00521-021-06392-x

[6] El-kenawy, E. S. M., Albalawi, F., Ward, S. A., Ghoneim, S. S., Eid, M. M., Abdelhamid, A. A., Bailek, N., & Ibrahim, A. (2022). Feature selection and classification of transformer faults based on novel meta-heuristic algorithm. *Mathematics*, *10*(17), 3144. https://doi.org/10.3390/math10173144

[7] Asteris, P. G., Douvika, M. G., Karamani, C. A., Skentou, A. D., Chlichlia, K., Cavaleri, L., Daras, T., Armaghani, D. J., & Zaoutis, T. E. (2020). A novel heuristic algorithm for the modeling and risk assessment of the COVID-19 pandemic phenomenon. *Computer Modeling in Engineering & Sciences*, *125*(2), 815-828. https://doi.org/10.32604/cmes.2020.013280

[8] Xue, J., & Shen, B. (2023). Dung beetle optimizer: A new meta-heuristic algorithm for global optimization. *The Journal of Supercomputing*, *79*(7), 7305-7336. https://doi.org/10.1007/s11227-022-04959-6

[9] Bourouis, S., Band, S. S., Mosavi, A., Agrawal, S., & Hamdi, M. (2022). Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images. *Frontiers in Oncology*, *12*, 834028. https://doi.org/10.3389/fonc.2022.834028

[10] Wen, W., Huang, Y., Xiao, Z., Tan, L., & Zhang, P. (2025). GAPSO: Cloud-edge-end collaborative task offloading based on genetic particle swarm optimization. *Symmetry*, *17*(8), 1225. https://doi.org/10.1007/s00607-021-00955-5

[11] Goli, A., Khademi-Zare, H., Tavakkoli-Moghaddam, R., Sadeghieh, A., Sasanian, M., & Malekalipour Kordestanizadeh, R. (2021). An integrated approach based on artificial intelligence and novel meta-heuristic algorithms to predict demand for dairy products: A case study. *Network: Computation in Neural Systems*, *32*(1), 1-35. https://doi.org/10.1080/0954898x.2020.1849841

[12] Goli, A., Ala, A., & Hajiaghaei-Keshteli, M. (2023). Efficient multi-objective meta-heuristic algorithms for energy-aware non-permutation flow-shop scheduling problem. *Expert Systems with Applications*, *213*, 119077. https://doi.org/10.1016/j.eswa.2022.119077

[13] Naruei, I., & Keynia, F. (2022). Wild horse optimizer: A new meta-heuristic algorithm for solving engineering optimization problems. *Engineering with Computers*, *38*(Suppl 4), 3025-3056. https://doi.org/10.1007/s00366-021-01438-z

[14] Myriam, H., Abdelhamid, A. A., El-Kenawy, E. S. M., Ibrahim, A., Eid, M. M., Jamjoom, M. M., & Khafaga, D. S. (2023). Advanced meta-heuristic algorithm based on Particle Swarm and Al-biruni earth radius optimization methods for oral cancer detection. *IEEE Access*, *11*, 23681-23700. https://doi.org/10.1109/access.2023.3253430

[15] Malik, A., Tikhamarine, Y., Souag-Gamane, D., Kisi, O., & Pham, Q. B. (2020). Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction. *Stochastic Environmental Research and Risk Assessment*, *34*(11), 1755-1773. https://doi.org/10.1007/s00477-020-01874-1

[16] Braik, M., Hammouri, A., Atwan, J., Al-Betar, M. A., & Awadallah, M. A. (2022). White shark optimizer: A novel bio-inspired meta-heuristic algorithm for global optimization problems. *Knowledge-Based Systems*, *243*, 108457. https://doi.org/10.1016/j.knosys.2022.108457

[17] Surakhi, O., Zaidan, M. A., Fung, P. L., Hossein Motlagh, N., Serhan, S., AlKhanafseh, M., Ghoniem, R. M., & Hussein, T. (2021). Time-lag selection for time-series forecasting using neural network and heuristic algorithm. *Electronics*, *10*(20), 2518. https://doi.org/10.3390/electronics10202518

[18] Yang, B., Wang, J., Zhang, X., Yu, T., Yao, W., Shu, H., Zeng, F., & Sun, L. (2020). Comprehensive overview of meta-heuristic algorithm applications on PV cell parameter identification. *Energy Conversion and Management*, *208*, 112595. https://doi.org/10.1016/j.enconman.2020.112595

[19] Momenitabar, M., Ebrahimi, Z. D., & Ghasemi, P. (2022). Designing a sustainable bioethanol supply chain network: A combination of machine learning and meta-heuristic algorithms. *Industrial Crops and Products*, *189*, 115848. https://doi.org/10.1016/j.indcrop.2022.115848

[20] Sangaiah, A. K., Hosseinabadi, A. A. R., Shareh, M. B., Bozorgi Rad, S. Y., Zolfagharian, A., & Chilamkurti, N. (2020). IoT resource allocation and optimization based on heuristic algorithm. *Sensors*, *20*(2), 539. https://doi.org/10.3390/s20020539

[21] Hafeez, G., Alimgeer, K. S., & Khan, I. (2020). Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid. *Applied Energy*, *269*, 114915. https://doi.org/10.1016/j.apenergy.2020.114915

[22] Seydanlou, P., Jolai, F., Tavakkoli-Moghaddam, R., & Fathollahi-Fard, A. M. (2022). A multi-objective optimization framework for a sustainable closed-loop supply chain network in the olive industry: Hybrid meta-heuristic algorithms. *Expert Systems with Applications*, *203*, 117566. https://doi.org/10.1016/j.eswa.2022.117566

[23] Tan, C., Hu, H., Ye, Q., Dianyu E., Cui, J., Zhou, Z., Kuang, S., Zou, R., & Yu, A. (2024). Multi-objective optimization of hydrocyclones using meta-heuristic algorithms and preference-informed decision-making. *Powder Technology*, *444*, 120050. https://doi.org/10.1016/j.powtec.2024.120050

[24] Naderi, E., Pourakbari-Kasmaei, M., Cerna, F. V., & Lehtonen, M. (2021). A novel hybrid self-adaptive heuristic algorithm to handle single-and multi-objective optimal power flow problems. *International Journal of Electrical Power & Energy Systems*, *125*, 106492. https://doi.org/10.1016/j.ijepes.2020.106492

[25] Ahmed, A. N., Van Lam, T., Hung, N. D., Van Thieu, N., Kisi, O., & El-Shafie, A. (2021). A comprehensive comparison of recent developed meta-heuristic algorithms for streamflow time series forecasting problem. *Applied Soft Computing*, *105*, 107282. https://doi.org/10.1016/j.asoc.2021.107282

[26] Khan, I. U., Javaid, N., Gamage, K. A., Taylor, C. J., Baig, S., & Ma, X. (2020). Heuristic algorithm based optimal power flow model incorporating stochastic renewable energy sources. *IEEE Access*, *8*, 148622-148643. https://doi.org/10.1109/access.2020.3015473

[27] Wang, J., Khishe, M., Kaveh, M., & Mohammadi, H. (2021). Binary chimp optimization algorithm (BChOA): A new binary meta-heuristic for solving optimization problems. *Cognitive*

Computation, *13*(5), 1297-1316. https://doi.org/10.1007/s12559-021-09933-7

[28] Jaaz, Z. A., Ezanee Bin Rusli, M., Rahmat, N. A., & Al-Adilee, M. K. A. (2025). Latency optimization approaches for healthcare Internet of Things and fog computing: A comprehensive review. *Journal of Intelligent Systems*, *34*(1), 20240543. https://doi.org/10.1515/jisys-2024-0543

[29] Zhang, J., Huang, Y., Ma, G., & Nener, B. (2021). Mixture optimization for environmental, economical and mechanical objectives in silica fume concrete: A novel frame-work based on machine learning and a new meta-heuristic algorithm. *Resources, Conservation and Recycling*, *167*, 105395. https://doi.org/10.1016/j.resconrec.2021.105395

[30] Abdalkareem, Z. A., Amir, A., Al-Betar, M. A., Ekhan, P., & Hammouri, A. I. (2021). Healthcare scheduling in optimization context: A review. *Health and Technology*, *11*(3), 445-469. https://doi.org/10.1007/s12553-021-00547-5

[31] Youn, S., Geismar, H. N., & Pinedo, M. (2022). Planning and scheduling in healthcare for better care coordination: Current understanding, trending topics, and future opportunities. *Production and Operations Management*, *31*(12), 4407-4423. https://doi.org/10.1111/poms.13867

[32] Wynendaele, H., Gemmel, P., Pattyn, E., Myny, D., & Trybou, J. (2021). Systematic review: What is the impact of self-scheduling on the patient, nurse and organization? *Journal of Advanced Nursing*, *77*(1), 47-82. https://doi.org/10.1111/jan.14579

[33] Zouari, F., Saad, K. B., & Benrejeb, M. (2012). Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. *International Review on Modelling and Simulations*, *5*(5), 2075-2103. https://doi.org/10.1109/ssd.2013.6564134

[34] Boulkroune, A., Hamel, S., Zouari, F., Boukabou, A., & Ibeas, A. (2017). Output-feedback controller based projective lag-synchronization of uncertain chaotic systems in the presence of input nonlinearities. *Mathematical Problems in Engineering*, *2017*(1), 8045803. https://doi.org/10.1155/2017/8045803

[35] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. *Results in Control and Optimization*, *11*, 100226. https://doi.org/10.1016/j.rico.2023.100226

[36] Boulkroune, A., Zouari, F., & Boubellouta, A. (2025). Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. *Journal of Vibration and Control*, 10775463251320258. https://doi.org/10.1177/10775463251320258

T-WGAN: A Transformer-Wasserstein GAN Approach for Melody Generation with Structural and Rhythmic Fidelity

Chunxiao Zhao
Zhengzhou Health College, Ministry of Health and Humanities Education (Art Education Center), Zhengzhou 450064, China
E-mail: zhaozhaoran1215@163.com

Keywords: music melody generation, deep neural network, transformer, generative adversarial network, rhythm consistency

Received: September 10, 2025

To address the current challenges of deep learning music generation models in capturing long-range dependencies, ensuring generation diversity, and maintaining training stability, this study proposes an optimized music melody generation model—T-WGAN. The model deeply integrates Transformer and Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP). This study first preprocesses the large-scale Lakh MIDI dataset, extracts single-track main melodies, and converts them into symbolic sequences using a REMI-based event representation. On this basis, the model innovatively adopts a generator based on Transformer decoder to learn the long-range structure of melodies. It also uses a critic based on Transformer encoder for stable adversarial training under the WGAN-GP framework to enhance the diversity and authenticity of generated melodies. Experimental results show that T-WGAN performs excellently on multiple key evaluation metrics. T-WGAN achieves a Rhythmic Consistency Rate (RCR) of 85.17%, significantly higher than baseline models (e.g., Transformer's 75.68%). Its score on Fréchet Distance for Music (FDM) drops to 31.02, proving that the generated melodies are closer to real music in feature distribution. The conclusion indicates that the proposed T-WGAN model successfully addresses the three core issues in melody generation—structural integrity, diversity, and training stability—synergistically. The findings provide an effective technical approach for generating high-quality music melodies with both structural logic and innovation.

Povzetek: Študija predstavlja model T-WGAN, ki izboljša generiranje glasbenih melodij z večjo raznolikostjo, boljšo strukturo in stabilnejšim učenjem.

1 Introduction

With the rapid development of artificial intelligence (AI) technology, its applications have expanded from traditional computing and analysis tasks to the field of artistic creation, which is a domain previously considered unique to human creativity. Among these applications, automatic music generation, as a typical representative of the intersection between AI and art, aims to create musical works with logic, aesthetic value, and emotional expressiveness through algorithms and models. It provides musicians with powerful auxiliary creation tools and opens new paradigms for content production in industries such as games, film and television scoring, and digital entertainment [1, 2]. However, music is not a simple accumulation of musical notes, but a time-series art that contains complex hierarchical structures, long-range dependencies, and subtle emotional connotations. Therefore, how to use deep neural network (DNN) to generate melodies with long-duration music with structural integrity, logical consistency, and originality has become a core scientific problem urgently needing to be solved in this field.

Amid the wave of deep learning, music generation technology has undergone evolution, from Recurrent Neural Network (RNN) and their variant, Long Short-Term Memory (LSTM) network, to various advanced generation models [3]. Early RNN/LSTM models, leveraging their ability to process sequential data, made significant progress in learning the local continuity of melodies. However, its inherent vanishing gradient problem and limitations in modeling long-range dependencies hinder its performance in generating structurally coherent long-duration music. To improve generation quality, researchers introduced models such as Generative Adversarial Network (GAN) and Variational Autoencoder (VAE). These models have shown great potential in enhancing the realism and diversity of generated samples [4]. Nevertheless, the training process of standard GAN is extremely unstable and often falls into the dilemma of mode collapse, resulting in monotonous and tedious generated melodies. In addition, although self-attention models represented by

Transformer can excellently capture long-range dependencies, when used independently for generation tasks, they still face challenges in training stability and generated sample diversity, making it difficult to effectively avoid the monotony of generated content. Therefore, how to integrate the advantages of different models and avoid their respective shortcomings constitutes the starting point of this study.

To address the above challenges, this study proposes an optimized music melody generation model based on the deep fusion of Transformer and Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP), which is called T-WGAN. The core innovation of this model lies in combining the strong global context capture capability of Transformer with the stable and high-quality generation training framework of WGAN-GP. This study aims to verify the following core hypotheses through experiments:

1) To evaluate the role of integrating Transformer as a generator in enhancing the coherence and accuracy of melodies in terms of long-range structure and rhythmic patterns.

2) To verify the effectiveness of the WGAN-GP training framework in overcoming the training instability of traditional GANs, avoiding mode collapse, and enhancing melody diversity, especially through comparison with the standalone Transformer model.

This study will conduct performance evaluation based on the large-scale Lakh MIDI dataset. The purposes are to explore an effective path for generating high-quality music melodies with both structural integrity and innovation, and to provide new insights and empirical support for the application of AI in the field of artistic creation.

2 Related work

2.1 Review of the research status of music generation

As reviewed in Wei et al., AI music technology was undergoing a transformation from an auxiliary "tool" to a "creator" with independent creative potential, and its application scenarios had been widely extended to fields such as digital entertainment, art education, and content creation [5]. The current research status presents a trend of high diversification and multimodal fusion. On one hand, research efforts continue to deepen in the sub-fields of music generation. For instance, Li et al. designed a generation method based on deep reinforcement learning for Xi'an Drum Music [6]. On the other hand, researchers are actively exploring the connection between music and other modal information, aiming to achieve "controllable generation" with stronger context awareness. For example, Zheng and Li realized real-time emotion-based piano music generation through GAN [7]. Huang et al. further extended the control conditions to the visual domain, using Creative-GAN to generate artistic music [8]. Kang et al. even utilized a multimodal Transformer model to successfully generate background music that aligns with the atmosphere of video content directly from the video itself [9].

2.2 Review of the research status of music generation based on DNN

DNN is the absolute core of current music generation research. Table 1 shows the research of related scholars.

Although Transformer performs excellently in structural modeling, when used independently for generation tasks, it still faces two key challenges: 1) the problem of training instability, which is likely to cause fluctuations in generation quality; 2) insufficient generation diversity, that is, the model may converge to some common and safe melody patterns, lacking innovation.

Table 1: Summary of the characteristics and limitations of music generation technology route based on DNN.

Technical route	Core ideas	Advantages	Main limitation	Representative literature
General framework	Sort out the representation, algorithms, evaluation methods, and challenges of symbolic music generation.	Provide a comprehensive theoretical framework and methodology in the field.	-	Ji et al. [10]
Adversarial network	Learn data distribution through the adversarial game between the generator and the discriminator.	Generate samples with high authenticity and the ability to learn complex distributions.	Unstable training, prone to mode collapse and poor diversity.	Liu [11], Tanawala and Dalwadi [12]
Attention mechanism	Use the self-attention mechanism to directly calculate the long-range	Effectively capture global structures and long-range dependencies.	Unstable training during independent generation, with a	Zhang et al. [13], Li and Sung [14], Wu et al. [15]

	relationships between elements in the sequence.		tendency to produce repetitive content.	
Multi-technology fusion	Combine deep learning with other technologies (e.g., genetic algorithms).	Enrich the technical paths of music generation and potentially solve specific problems.	Increased model complexity and training difficulty.	Majidi and Toroghi [16]

2.3 Research gap and innovation of this study

Through a review of existing DNN-based music generation technical routes, it is evident that although various models have achieved significant progress in specific aspects, the current research field lacks an integrated solution that can synergistically address the three core challenges: long-range structural dependencies, generation diversity, and training stability. The proposed T-WGAN model aims to fill this critical gap. Its core innovation lies in the strong integration of Transformer into the stable and efficient adversarial training framework of WGAN-GP, with the expectation of providing a reliable reference direction for the generation of high-quality music melodies.

3 The design and evaluation method of melody generation model

3.1 The overall framework design analysis of melody generation model

To effectively address the key issues in music melody generation, including insufficient modeling of long-range dependencies, lack of diversity in generated melodies, and instability in adversarial training, this study proposes the T-WGAN model, which integrates Transformer with the WGAN-GP mechanism [17]. The overall framework of the T-WGAN model is shown in Fig. 1.

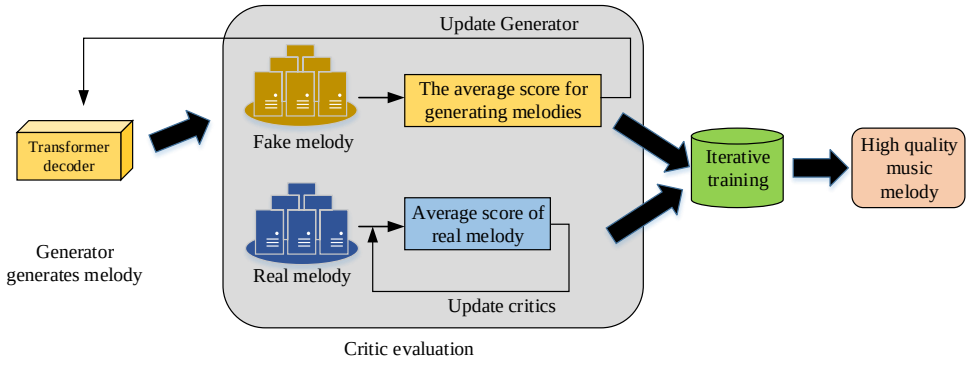


Figure 1: Schematic diagram of melody generation model architecture based on T-WGAN.

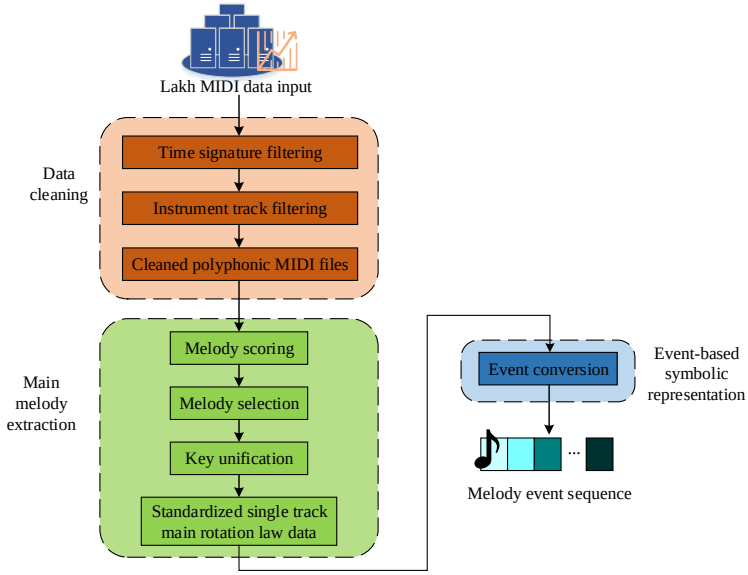


Figure 2: Schematic diagram of pretreatment process of training data.

3.2 Training data preprocessing and symbolic representation

This study adopts The Lakh MIDI Dataset v0.1 (LMD) (https://colinraffel.com/projects/lmd/) as the original data source for the melody generation task. This section focuses on a three-step process of "data cleaning-melody extraction-symbolic representation" (Fig. 2), and finally obtains high-quality single-melody sequence data suitable for T-WGAN model training.

In the data cleaning stage, first, music pieces are filtered based on time signature information, retaining only samples with the common 4/4-time signature to avoid rhythm alignment difficulties caused by complex time signatures. Then, MIDI tracks are classified by instrument, only instruments with strong melodic properties (such as piano, flute, and violin) are retained, while percussion instruments and chord filling tracks are filtered out. Next, the "main melody extraction algorithm" is used to select the track with the strongest melodic feature. Specifically, a heuristic scoring function based on pitch range and track activity is adopted to score all non-percussion tracks of each piece of music. The track with the highest score is selected as the main melody track, as shown in equation (1):

$$S_i = \alpha \cdot \text{PitchRange}(i) + \beta \cdot \text{NoteDensity}(i) \quad (1)$$

S_i denotes the melody score of the i-th track. $\text{PitchRange}(i)$ is the pitch range of this track. NoteDensity(i) is the number of notes per unit time. The empirical weights α and β are set to 0.6 and 0.4 respectively, aiming to balance the preference for melodies with wider melodic contours and richer rhythmic activities. In addition, to eliminate the burden on the model caused by different keys, all melodies are transposed to C major or a minor. The original key is identified using a key estimation tool (such as the key.estimate()) function provided by the music21 library) and converted through a transposition function to ensure the consistency and generalization ability of the input space.

To enhance the model's ability to understand music structure, this study adopts an event-based symbolic method to encode melody data, replacing the fixed-time grid representation of traditional Piano Roll. For example, a note event representing "playing a note with Pitch 60, Duration of 1/8 beat, and Velocity 80 at the 3rd Position of the 2nd Bar" can be encoded into the event sequence shown in equation (2):

$$\text{Bar}_2 \rightarrow \text{Position}_3 \rightarrow \text{Pitch}_{60} \rightarrow \text{Duration}_{1/8} \rightarrow \text{Velocity}_{80} \quad (2)$$

This encoding method not only preserves the structural information of notes but also effectively models the interactive relationships across dimensions such as rhythm, dynamics, and time, adapting to the Transformer structure's requirement for discrete sequence modeling. For example, the event dictionary maps MIDI pitches to tokens from Pitch_21 to Pitch_108, and quantized positions based on sixteenth notes to tokens from Pos_0 to Pos_15. By analogy, a vocabulary containing approximately 300 independent events is constructed. For rare events not covered by the dictionary (such as irregular time signatures or microtones), the preprocessing pipeline filters them out to maintain the consistency of the symbolic space. Overlapping notes (polyphony) appearing in the main melody track are processed by prioritizing the highest pitch at any time step, thereby ensuring the generation of monophonic melody sequences suitable for the focus of this study.

3.3 Optimization strategy analysis of melody generation model based on T-WGAN

The proposed T-WGAN model deeply integrates the Transformer architecture [18] with the optimization mechanism of WGAN-GP, constructing a music generation network capable of capturing melodic structures, ensuring training stability, and enhancing generation diversity. Its main components include a generator built based on the Transformer decoder and a critic built based on the Transformer encoder. These two components form a typical adversarial learning architecture as shown in Figure 3. The specific architectural details of the generator and the critic are presented in Table 2. Both the generator and the critic adopt standard positional encoding to inject sequence order information, and apply layer normalization within each Transformer block, which is consistent with the original Transformer architecture.

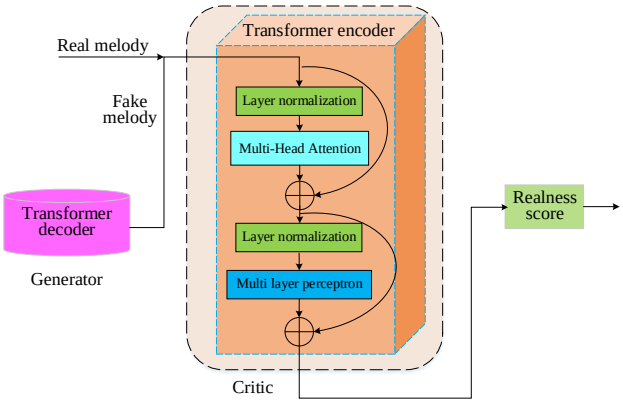


Figure 3: Schematic diagram of antagonistic learning architecture of melody generation model based on T-WGAN.

Table 2: Architecture detail table of T-WGAN model components

Parameter	Generator (Transformer decoder)	Critics (Transformer encoder)
Embedding dimension	256	256
Number of layers	6	6
Attention heads	8	8
Feed-forward dim	1024	1024
Activation function	ReLU	LeakyReLU
Dropout rate	0.1	0.1

In this architecture, the task of the generator is to learn the mapping function G(z) from the random noise z∼N(0,I) in the potential space to the melody event space, and gradually generate the melody sequence by autoregressive method, as shown in equation (3):

$$x = G(z) = (x_1, x_2, \ldots, x_T) \tag{3}$$

$x_t \in V$ represents the event generated at the t time step (such as Pitch64, Duration1/8, etc.), and V is the event dictionary. Specifically, the generator predicts that the event probability distribution of each time step is equation (4):

$$P(x_t | x_{<t}, z) = Softmax(f_\theta(x_{<t}, z)) \tag{4}$$

f_θ represents the output function modeled by the Transformer decoder, and the parameter is θ. The generation process of the whole sequence is to maximize conditional likelihood, as shown in equation (5):

$$L_{MLE} = \sum_{t=1}^{T} logP(x_t | x_{<t}, z) \tag{5}$$

This autoregressive modeling method relies on Transformer's Multi-Head Attention mechanism, which can effectively capture the long-term dependence in melody structure.

In confrontation training, the Critic's goal is to learn a real function $D(x) \in R$, which is used to measure whether the input melody is close to the real data. Different from the traditional binary classifier in GAN, WGAN introduces the Wasserstein-1 distance between the real sample distribution P_r and the generated sample distribution P_g as the optimization objective, which is defined as equation (6):

$$W(P_r, P_g) = \sup_{\|f\|_L \le 1} E_{x \sim P_r}[f(x)] - E_{\tilde{x} \sim P_g}[f(\tilde{x})] \tag{6}$$

f is a 1-Lipschitz continuous function, and Critic D(x)≈f(x). The objective is to maximize this expected difference, such that real samples receive higher scores and generated samples receive lower scores. To ensure the Critic satisfies the 1-Lipschitz constraint, WGAN-GP replaces the previous weight clipping method by introducing a GP term. The final objective function is shown in equation (7):

$$L_D = E_{\tilde{x} \sim P_g}[D(\tilde{x})] - E_{x \sim P_r}[D(x)] + \lambda E_{\hat{x} \sim P_{\hat{x}}}\left(\left\|\nabla_{\hat{x}} D(\hat{x})\right\|_2 - 1\right)^2 \tag{7}$$

λ refers to the GP coefficient, which is taken as λ = 10 in this study. $\hat{x}$ is an intermediate sample obtained by linear interpolation between the real sample $x \sim P_r$ and the generated sample $\tilde{x} \sim P_g$, as shown in equation (8):

$$\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}, \quad \varepsilon \sim U[0,1] \tag{8}$$

The goal of the Critic's update is to minimize L_G, while the optimization goal of the generator is to maximize the Critic's score on the generated samples, as shown in equation (9):

$$L_G = -E_{\tilde{x} \sim P_g}[D(\tilde{x})] \tag{9}$$

In this way, and generators constitute a "zero-sum game" optimization process, which is expressed as $\min_G \max_D L_D$. During the adversarial training process, this study adopts an update ratio of 5:1 between the Critic and the Generator—i.e., the Critic is updated five times for each update of the Generator. While the optimal update ratio may vary depending on specific tasks and model architectures, a low ratio (e.g., 1:1) may lead to insufficient training of the Critic. This results in unstable gradient directions provided by the Critic, thereby causing training collapse. An excessively high ratio, however, will unnecessarily increase computational costs. Therefore, the adoption of the fully validated 5:1 ratio is a reliable compromise between ensuring gradient quality and maintaining reasonable training efficiency. It is particularly suitable for symbolic sequence generation tasks that require refined gradient guidance, effectively avoiding mode collapse and promoting diversity.

Algorithm 1: T-WGAN Training Procedure

Require: Generator G with parameters θg, Critic C with parameters θc, real data distribution Pr, noise dimension dz, batch size m, Critic update iterations n_{critic}, GP coefficient λ.

Require: Adam optimizers opt$_g$, opt$_c$.

while not converged **do**

  for t=1,…,n_{critic} **do**

    Sample a batch of real melodies $\{x^{(i)}\}_{i=1}^{m} \sim Pr$

    Sample a batch of noise vectors $\{z^{(i)}\}_{i=1}^{m} \sim N(0, I)$

    Generate a batch of fake melodies

    Sample $\varepsilon \sim U[0,1]$

    Create interpolated samples $\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}$

    $C_{\text{real}} = C(x)$

    $C_{\text{fake}} = C(\tilde{x})$

    // Calculate Gradient Penalty

    $g = \nabla_{\hat{x}} C(\hat{x})$

    $L_{GP} = \lambda \cdot (\| g \|_2 - 1)^2$

    // Critic Loss

    $L_C = E[C_{fake}] - E[C_{real}] + E[L_{GP}]$

    // Update Critic

    Zero gradients of opt$_c$.

    L_C.backward()

    opt$_c$.step()

  end for

  // Update Generator

  Sample a batch of noise vectors $\{z^{(i)}\}_{i=1}^{m} \sim N(0, I)$ ·

28: Generate a batch of fake melodies $\tilde{x} = G(z)$
29:
30: // Generator Loss
31: $L_G = -E[C(\tilde{x})]$
32:
33: // Update Generator
34: Zero gradients of opt_g
35: L_G .backward()
36: opt_g. step()
37: **end while**

Among them, the pseudocode flow of the melody generation model based on T-WGAN constructed in this study is shown in algorithm 1.

3.4 Experimental evaluation

In this study, the Lakh MIDI Dataset v0.1 is selected as the experimental dataset, which contains more than 170,000 music works in MIDI format, covering various styles such as pop, jazz, and classical. To ensure training quality, the data is cleaned and filtered, mainly including: 1) Removing samples with a duration of less than 30 seconds or fewer than 200 events. 2) Standardizing the tempo of all MIDI tracks (unified to 120 BPM). 3) Retaining only the main melody track and extracting its pitch, duration, and velocity as model inputs. Finally, approximately 38,000 high-quality single-track melody segments are organized for model training and validation, with the ratio of training set, validation set, and test set being 8:1:1.

The specific software and hardware configurations as well as hyperparameter settings are presented in Table 3. These parameter settings combine the common configurations of Transformer models in the field of music generation and have been fine-tuned through preliminary experiments to balance model performance and computational efficiency on the Lakh MIDI dataset. For example, 6 network layers and 8 attention heads represent a compromise between capturing long-range dependencies and preventing overfitting. In the symbolic representation, the MIDI timeline is quantized to a resolution of sixteenth notes. Each discrete step in the generated sequence, referred to as a 'Time Step', corresponds to such a time quantum. Therefore, in a standard 4/4-time piece, one measure consists of 16-time steps.

In the model evaluation phase, the proposed T-WGAN model algorithm is compared with Transformer, MuseGAN [19], MusicVAE [20], Pop Music Accompaniment Generator (PopMAG) [21], and the model algorithm proposed by Wu et al. (2024) in related fields. The following metrics are used to measure performance: Note Repetition Rate (NRR), Mean Pitch Error (MPE), Rhythmic Consistency Rate (RCR), and Fréchet Distance for Music (FDM).

These metrics are selected in this study to comprehensively evaluate the quality of generated melodies from multiple dimensions: NRR is used to measure melody diversity and avoid monotony and repetition. MPE calculates the mean absolute error between the pitch predicted by the model and the real pitch (both represented by MIDI pitch numbers) at each time step. RCR examines the coherence of rhythmic structures. FDM evaluates similarity by first encoding generated and real melody segments into a pre-trained music feature embedding space (features are usually extracted using audio or symbolic music models such as VGGish), then calculating the Fréchet distance between these two sets of feature distributions. These metrics together form a multi-dimensional evaluation system.

Table 3: Software and hardware configuration and hyperparameter setting.

Category	Parameter term	Setting value/description
Data preprocessing	Minimum number of events	≥ 200
	Minimum duration	≥ 30 seconds
	Rhythm unification	120 BPM
	Input feature	Pitch, Duration, Velocity
Software environment	Framework	PyTorch 2.1
	Operating system	Ubuntu 22.04
Hardware environment	GPU	NVIDIA A100 80GB
	CPU	Intel Xeon Platinum
Hyperparameter	Learning rate	0.0005
	Batch Size	64
	Maximum training rounds	100
	Optimizer	Adam
	Loss function	Cross-Entropy
	Stop strategy	Early Stopping (Verification set 5 rounds without improvement termination)

4 Result and discussion

4.1 Pitch and diversity analysis under different algorithms

Comparing this model algorithm with other algorithms, the results of pitch and diversity generated by music melody are shown in Fig. 4 and Fig. 5.

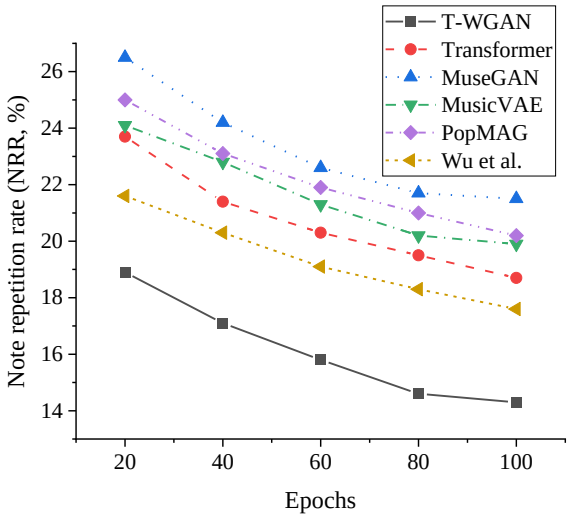


Figure 4: Results of note repetition rate generated by music melody under different algorithms.

In Fig. 4, throughout different training epochs, the proposed T-WGAN model consistently outperforms all comparative models in the NRR metric, demonstrating a stronger ability to generate melodic diversity. In the early stage of training (20 epochs), the NRR of T-WGAN is 18.9%, which is already significantly lower than that of Transformer (23.7%) and MuseGAN (26.5%). As training progresses, the NRR of T-WGAN continues to decrease steadily to 14.3%, outperforming the second-best model proposed by Wu et al. (17.6%). This advantage can be attributed to the introduction of the WGAN-GP training framework. By optimizing the Wasserstein distance, this framework effectively alleviates the mode collapse problem, prompting the generator to explore a broader melodic space. This avoids the monotonous and repetitive patterns that standalone Transformer models tend to fall into due to error accumulation in autoregressive generation.

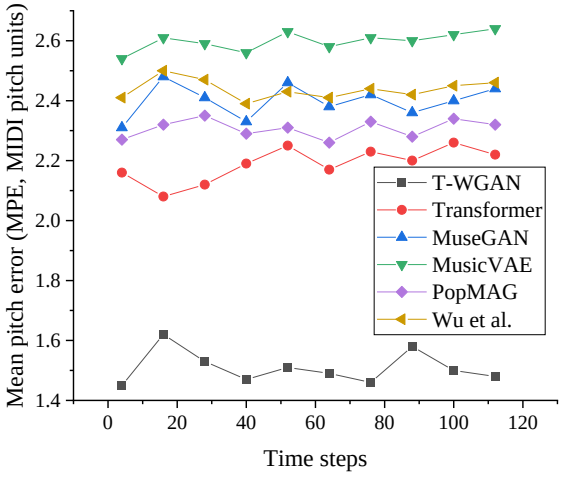


Figure 5: Results of average pitch error generated by music melody under different algorithms.

Meanwhile, data in Fig. 5 shows that the T-WGAN model maintains a significantly lower MPE than comparative models across all time steps, demonstrating excellent accuracy in melodic pitch prediction. In short-range contexts (e.g., Time step=4), the MPE of T-WGAN is 1.45, far lower than that of Transformer (2.16), MuseGAN (2.31), and MusicVAE (2.54). Even in long-range contexts (e.g., Time step=112), the error value of T-WGAN remains at a low level of 1.48, while the errors of other models are generally higher than 2.3. This is mainly due to the Transformer architecture serving as the generator backbone. Its self-attention mechanism can effectively capture long-range dependencies in music sequences, enabling the model to consider broader contextual information when predicting the current pitch, thus ensuring the coherence of the melody in global structure and the accuracy of local pitches.

4.2 Evaluation and analysis of music melody generation effect under different algorithms

The proposed model algorithm is compared with other algorithms, and the results of rhythm consistency and distribution similarity of music melody generation are further evaluated, as shown in Figs. 6 and 7.

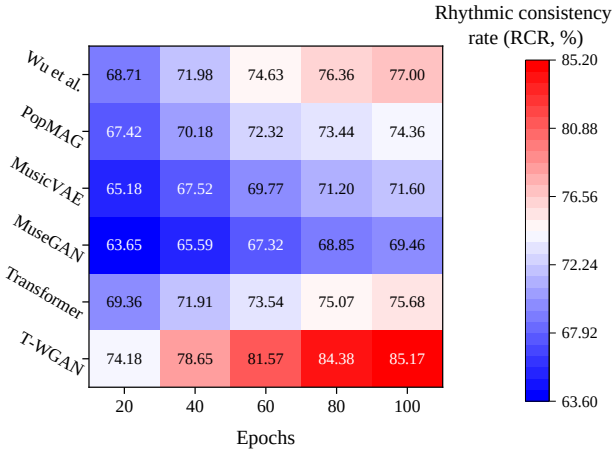


Figure 6: Results of rhythm retention rate of music melody generation under different algorithms.

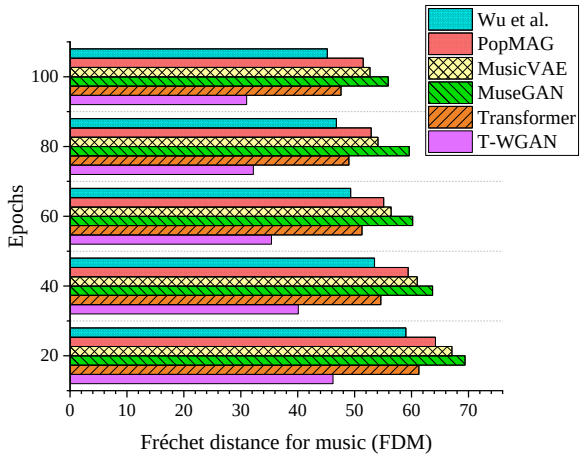


Figure 7: Results of generation fragment similarity of music melody generation under different algorithms.

In Figs. 6 and 7, the T-WGAN model demonstrates significant advantages in two core metrics: rhythmic consistency and musical fidelity. Regarding RCR, T-WGAN maintains a leading position across all training epochs. Its RCR steadily increases from 74.18% at the 20th epoch to 85.17% at the 100th epoch, significantly outperforming Transformer (75.68%), MuseGAN (69.46%), and the method proposed by Wu et al. (77.00%). As for the FDM metric, it comprehensively evaluates musical fidelity by measuring the distribution distance between generated samples and real samples in the feature space. The FDM score of T-WGAN steadily decreases from an initial 46.20 to 31.02, which is significantly better than all comparative models, such as Transformer (47.60) and MuseGAN (55.90). The substantial reduction in the FDM score holds important musical significance: it indicates that the melodies generated by T-WGAN are closer to real music in the joint distribution of multi-dimensional features, including pitch, rhythm, and dynamics.

To further verify the reliability of the performance improvement of the T-WGAN model, paired t-tests are conducted between its final results (after 100 training epochs) on the four core evaluation metrics and those of all baseline models. The test results are presented in Table 4, with all p-values evaluated at a significance level of $\alpha = 0.05$. The results indicate that the superiority of T-WGAN is not merely due to random fluctuations or accidental outcomes but is statistically significant. This finding strongly supports the conclusion of this study: the T-WGAN model significantly outperforms existing methods in overall performance.

Table 4: Paired t-test results of T-WGAN and baseline model on each index.

Contrast model	NRR (%)	MPE	RCR (%)	FDM
Transformer	0.002	< 0.001	0.001	< 0.001
MuseGAN	< 0.001	< 0.001	< 0.001	< 0.001
MusicVAE	< 0.001	< 0.001	< 0.001	< 0.001
PopMAG (MuseFlow)	0.003	< 0.001	< 0.001	< 0.001
Wu et al. [15]	0.009	< 0.001	< 0.001	< 0.001

4.3 Ablation experiment

To systematically evaluate the individual contributions of the Transformer architecture and the WGAN-GP training framework in the T-WGAN model, the following ablation study is designed. Quantitative analysis is conducted on the expected performance of each variant model, as detailed in Table 5.

Table 5: Variant model of ablation research and numerical analysis of expected performance.

Model variant	NRR (%)	MPE	RCR (%)	FDM
1. Transformer-only	23.7	2.16	75.68	47.6
2. Transformer + vanilla GAN	~20.0 (unstable)	> 3.0 (bad)	~65.0 (bad)	> 60.0 (bad)
3. WGAN-GP + RNN Generator	15.5	2.55	68.2	50.5
4. T-WGAN (the proposed model)	14.3	1.48	85.17	31.02

In Table 5, the ablation analysis clearly demonstrates the complementarity and necessity of each component in the T-WGAN model. The Transformer-only baseline performs well in the structural metric (RCR) but underperforms in diversity (NRR) and fidelity (FDM). The introduction of a standard GAN (Transformer + vanilla GAN) may lead to an overall deterioration of all performance metrics due to extremely unstable training. On the other hand, using the WGAN-GP framework alone with a weaker RNN generator (WGAN-GP + RNN Generator) can ensure generation diversity (low NRR) but sacrifices the ability to capture long-range musical structures (low RCR). Therefore, only the combination of Transformer and WGAN-GP can achieve synergistic improvements across all key dimensions, thereby generating high-quality melodies that possess both structural integrity and content novelty.

4.4 Discussion

The experimental results of this study indicate that the success of T-WGAN stems from the synergistic advantages of its architecture. As the generator backbone, the Transformer ensures the coherence of melodies in long-range structures, which is consistent with the finding of Agarwal and Sultanova [22] that emphasizes the core role of the Transformer. Meanwhile, the WGAN-GP training framework effectively overcomes mode collapse and significantly enhances generation diversity. The high RCR and low MPE are manifestations of the model's mastery of musical "grammar" (i.e., structural rules), as it can generate long sequences with coherent rhythms and accurate pitches. This suggests the model is not merely imitating notes but learning their inherent structural relationships. Compared with the method of improving stability through ensemble learning proposed by Nag et al. [23], this deep integration of "structure awareness" and "stable generation" achieves more efficient performance improvement within a single model. Nevertheless, the model still has limitations. For instance, this study relies entirely on objective metrics and lacks human subjective evaluation of the aesthetic value of generated music. As investigated by Dong [24], future work can explore the use of heuristic methods such as genetic algorithms for post-processing optimization of

generated melodies, and must introduce subjective listening experiments to construct a more comprehensive evaluation system.

5 Conclusion

This study successfully constructs a deeply integrated model based on T-WGAN, which combines the powerful long-range dependency capture capability of Transformer with the stable training framework of WGAN-GP. Experimental results show that T-WGAN significantly outperforms existing baseline models across multiple objective metrics, such as note diversity, pitch accuracy, rhythmic consistency, and generation fidelity. With a pitch error of only 1.45, it demonstrates advancement in generating high-quality, structurally complete, and innovative melodies.

Nevertheless, this study still has clear limitations. First, the model currently focuses on monophonic melody generation and has not been extended to more complex polyphonic music creation such as harmony and orchestration. Second, the evaluation system of this study relies entirely on objective metrics, lacking subjective evaluations by human listeners on the aesthetic value and emotional expression of generated music—this is a key link in measuring the success of music creation. Third, the study fails to provide detailed ablation experiments to accurately quantify the individual contributions of the Transformer and WGAN-GP to the final performance.

Looking forward, future research can be deepened in the following three aspects: 1) Extend the T-WGAN framework to multi-track and polyphonic music generation tasks; 2) Introduce conditional variables such as style and emotion to explore controllable music generation, thereby enhancing the model's practicality and interactivity; 3) Combine subjective listening experiments with objective metrics to construct a more comprehensive evaluation system for music generation quality to further promote the application of AI in the field of music creation.

References

[1] Sadiku M N, Ajayi S A, Sadiku J O. Artificial intelligence in media and entertainment. International Journal of Scientific and Applied Research (IJSAR), 2025, 5(5): 1-4. https://doi.org/10.54756/IJSAR.2025.5.1

[2] Kwiecień J, Skrzyński P, Chmiel W, Dąbrowski A, Szadkowski B, Pluta M. Technical, musical, and legal aspects of an ai-aided algorithmic music production system. Applied Sciences, 2024, 14(9): 3541. https://doi.org/10.3390/app14093541

[3] Kumar J, Goomer R, Singh AK. Long short-term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters. Procedia Computer Science, 2018, 125: 676-682. https://doi.org/10.1016/j.procs.2017.12.087

[4] Dash A, Agres K. AI-based affective music generation systems: A review of methods and challenges. ACM Computing Surveys, 2024, 56(11): 287. https://doi.org/10.1145/3672554

[5] Wei L, Yu Y, Qin Y, Zhang S. From tools to creators: A review on the development and application of artificial intelligence music generation. Information, 2025, 16(8): 656. https://doi.org/10.3390/info16080656

[6] Li P, Liang T, Cao Y, Wang X, Wu X, Lei L. A novel Xi'an drum music generation method based on Bi-LSTM deep reinforcement learning. Applied Intelligence, 2024, 54(1): 80-94. https://doi.org/10.1007/s10489-023-05195-y

[7] Zheng L, Li C. Real-time emotion-based piano music generation using generative adversarial network (GAN). IEEE Access, 2024, 12: 87489-87500. https://doi.org/10.1109/ACCESS.2024.3414673

[8] Huang J, Huang X, Yang L, Tao Z. Dance-conditioned artistic music generation by creative-GAN. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2024, E107.A (5): 836-844. https://doi.org/10.1587/transfun.2023EAP1059

[9] Kang J, Poria S, Herremans D. Video 2 music: Suitable music generation from videos using an affective multimodal transformer model. Expert Systems with Applications, 2024, 249: 123640. https://doi.org/10.1016/j.eswa.2024.123640

[10] Ji S, Yang X, Luo J. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. ACM Computing Surveys, 2023, 56(1): 7. https://doi.org/10.1145/3597493

[11] Liu W. Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition. The Journal of Supercomputing, 2023, 79(6): 6560-6582. https://doi.org/10.1007/s11227-022-04914-5

[12] Tanawala B A, Dalwadi D C. Harmonic synergy: Leveraging deep convolutional networks, LSTMs, and RNNs for multi-genre piano roll generation with GANs. SN Computer Science, 2025, 6(4): 297. https://doi.org/10.1007/s42979-025-03855-z

[13] Zhang Y, Zhou Y, Lv X, Li J, Lu H, Su Y, Yang H. TARREAN: a novel transformer with a gate recurrent unit for stylized music generation. Sensors, 2025, 25(2): 386. https://doi.org/10.3390/s25020386

[14] Li S, Sung Y. MRBERT: pre-training of melody and rhythm for automatic music generation. Mathematics, 2023, 11(4): 798. https://doi.org/10.3390/math11040798

[15] Wu S L, Donahue C, Watanabe S, Bryan N J. Music controlnet: Multiple time-varying controls for music generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2692-2703. https://doi.org/10.1109/TASLP.2024.3399026

[16] Majidi M, Toroghi R M. A combination of multi-objective genetic algorithm and deep learning for

music harmony generation. Multimedia tools and applications, 2023, 82(2): 2419-2435. https://doi.org/10.1007/s11042-022-13329-6

[17] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A C. Improved training of Wasserstein Gans. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach California USA, December 4-9, 2017, pp. 5769-5779. https://dl.acm.org/doi/10.5555/3295222.3295327

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach California USA, December 4-9, 2017, pp. 6000-6010. https://dl.acm.org/doi/10.5555/3295222.3295349

[19] Wang F. Application of artificial intelligence-based music generation technology in popular music production. Journal of Combinatorial Mathematics and Combinatorial Computing, 2025, 127: 655-671. https://doi.org/10.61091/jcmcc127a-038

[20] Tie Y, Guo X, Zhang D, Tie J, Qi L, Lu Y. Hybrid learning module-based transformer for multitrack music generation with music theory. IEEE Transactions on Computational Social Systems, 2024, 12(2), 862-872. https://doi.org/10.1109/TCSS.2024.3486604

[21] Zhu H, Liu Q, Yuan NJ, Zhang K, Zhou G, Chen E. Pop music generation: From melody to multi-style arrangement. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(5): 54. https://doi.org/10.1145/3374915

[22] Agarwal S, Sultanova N. Music generation through transformers. International Journal of Data Science and Advanced Analytics, 2024, 6(6): 302-306. https://doi.org/10.69511/ijdsaa.v6i6.231

[23] Nag B, Middya A I, Roy S. Melody generation based on deep ensemble learning using varying temporal context length. Multimedia Tools and Applications, 2024, 83(27): 69647-69668. https://doi.org/10.1007/s11042-024-18270-4

[24] Dong L. Using deep learning and genetic algorithms for melody generation and optimization in music. Soft Computing, 2023, 27(22): 17419-17433. https://doi.org/10.1007/s00500-023-09135-3

Enhanced Prediction of Manufacturing Quality Ratings Using Optimized Stacking Ensemble Modeling

Lanbao Hou, Zhiqiong Zou*
School of Mathematics and Physics, Jingchu University of Technology, Jingmen 448000, Hubei, China
*Corresponding author
E-mail: 11095881@163.com

Keywords: manufacturing quality, machine learning (ML), stacking, SHAP analysis, material transformation metric (MTM)

Received: May 20, 2025

The predictive modeling of quality in manufacturing is vital in improving efficiency, cutting costs, and attaining zero-defect production. This paper addresses this requirement through an empirical comparison of machine learning models to predict manufacturing quality ratings. On a simulated dataset of 3,957 samples with five important features of the process, this study compared the performance of six basic models (AdaBoost, Bagging, Decision Tree, Gradient Boosting, K-Nearest Neighbors, and XGBoost) and two more advanced ensemble models (Averaging and Stacking). The findings demonstrated that the Stacking Ensemble model outperformed all other candidates, with higher performance in terms of an R^2 of 0.999 and the lowest values of error (MSE: 0.004, RMSE: 0.065, MAE: 0.016) in the test set. Moreover, SHAP analysis of the Stacking ensemble as the best model revealed that Material Transformation Metric (MTM) and Temperature (T) were the top features having a significant impact on the quality of products. The analysis finds that the Stacking Ensemble method provides a valid and very effective framework for predicting quality, which is helpful in the optimization of the manufacturing process.

Povzetek: Članek primerja več modelov strojnega učenja za napoved ocene kakovosti v proizvodnji in pokaže, da je ansambelski model stacking *najboljši (R^2≈0,999, najnižje napake), pri čemer sta po analizi SHAP najpomembnejša vplivna dejavnika MTM in temperatura.*

1 Introduction

Manufacturing has evolved significantly with the advent of the new sophisticated information systems into smart manufacturing. Quality of products in this extremely competitive environment will be a decisive factor in profitability, competitiveness and sustainability. On the other hand, low quality results in higher expenses, loss of reputation and customers. High-quality and zero-defect manufacturing production necessitates strong systems to anticipate and manage quality variables in advance.

In this regard, machine learning (ML) has become a potent solution to examine large volumes of manufacturing data to forecast quality results. ML models can compute the complex trends and relationships between the parameters of processes and final product quality, using historical data, thus making it possible to optimize and model manufacturing processes [1]. Proper quality prediction gives huge benefits throughout the supply chain, allowing correction beforehand and resource conservation [2].

The application of ML in this area has been extensively studied. In order to give a systematic description of the existing research environment, Table 1 describes the key related works, with their diversity in terms of dataset, methodology, and contribution.

Table 1: Summary of related works in manufacturing quality prediction using ML.

Reference	Dataset & Context	Methodology	Key Contributions	Main Findings
Li et al. [1]	Manufacturing quality data	Combined system: Particle Swarm Optimizer + Neural Networks	Proposed an intelligent prediction model and evaluation system based on big data.	The model's performance aligned with actual manufacturing outcomes, proving suitable for real-world construction activities.
Sankhye and Hu [2]	Product quality data	Not Specified (Focus on feature construction)	Emphasized the critical role of feature engineering and expert knowledge in training	Highlighted that prior quality understanding can save costs related to recalls, packaging, and transferring.

			models for quality prediction.	
Link et al. [3]	Manufacturing quality data	Integration of expert knowledge with ML models	Combined heuristic expert knowledge with data-driven ML to predict product quality.	Direct expert involvement saved time and aided model interpretation, while ML compensated for human iteration limits.
Weichert et al. [4]	Review article (Various manufacturing processes)	Review of ML and optimization methods	Provided a comprehensive review of ML for optimizing production processes.	Stressed the need to consider the correlation between data, algorithms, optimizers, and the specific quality problem.
Psarommatis and Azamfirei [5]	Conceptual/Review	Review and framework for Zero Defect Manufacturing (ZDM)	Provided a complete guide for advanced and sustainable quality management towards ZDM.	Positioned ZDM as a holistic approach requiring a comprehensive examination of the entire manufacturing system.
Msakni et al. [6]	Automotive product quality	Neural Networks, Random Forest, LSTM	Applied ML to forecast specific automotive product quality and detect tolerance violations.	Enabled process improvements by early detection of quality issues, preventing costly defects.
Yang et al. [7]	Manufacturing firm data	Dual Machine Learning Models	Investigated the relationship between ESG (Environmental, Social, Governance) ratings and digital technological innovation.	Showed that external factors like governance and society can influence manufacturing technology adoption.
Akbari et al. [8]	Spatter behavior in Laser Powder Bed Fusion (LPBF)	Six ML methods (Neural Networks outperformed others)	Predicted ejection velocity and spatter direction to minimize defects in additive manufacturing.	Neural Networks outperformed other models, significantly progressing defect identification and product quality improvement.
Azamfirei et al. [9]	Manufacturing quality inspection	Automation and in-line inspection systems	Applied automation for in-line quality inspection as a zero-defect manufacturing approach.	Demonstrated the practical application of automated systems to achieve high-quality production goals.
This Study	Simulated manufacturing process data (3,957 samples, 5 features: T, P, TxP, MFM, MTM)	**Comprehensive comparison of 8 models: AdaBoost, Bagging, DT, GB, KNN, XGBoost, Averaging, Stacking**	**Extensively compared multiple base and ensemble models; Identified Stacking as optimal; Used SHAP for granular feature importance analysis.**	**Stacking Ensemble achieved superior performance (R^2: 0.999, lowest errors); Quantified feature influence (MTM most critical, P least).**

Even as the literature indicates that ML has enormous potential in quality prediction, there still exist gaps. Most of the literature is based on a few models or localized specific manufacturing settings, which limits the generalizability of their results [2], [6]. Moreover, though ensemble methods are known to be powerful, more extensive comparisons of a broad range of base and ensemble models, namely with respect to the manufacturing quality rating, are less frequent. Additionally, model interpretability is also desired to understand which process parameters have the strongest effect on quality outcome, which is a feature that is essential to practical implementation. This study will address these gaps by conducting a general empirical comparison of different emerging ML models, including base learners and state-of-the-art ensemble models, in predicting manufacturing quality ratings. Furthermore, this study goes beyond model execution to offer practical information using explainable AI (XAI) methods, i.e., SHAP analysis, to prioritize the effect of essential manufacturing characteristics.

The ultimate objective of the paper is to construct, analyze and assess a predictive model of quality rating in manufacturing. The given goals are achieved by the following specific research questions:

1) Which of the following ML models (AdaBoost, Bagging, Decision Tree, Gradient Boosting, K-Nearest Neighbors, and XGBoost) proves to be the most accurate and efficient in forecasting manufacturing quality rating?
2) Can more sophisticated ensemble methods (Averaging and Stacking) increase prediction performance than the best single model?
3) How do the relative values of the major process parameters, Temperature (T), Pressure (P), their combination (TxP), Material Fusion Metric (MFM), and Material Transformation Metric (MTM) affect the predicted quality rating?

The study hypothesizes that the Stacking Ensemble model will outperform all the base models since it can be used to combine their strengths and thus give the most effective and reliable results in predicting manufacturing quality.

The remainder of the paper will consist of the following: Section 2 will contain the description of the methodology, including the dataset, computational environment and the metrics employed. Section 3 discusses the ML models, and hyperparameter optimization is then explained in Section 4. The results and findings are given in Section 5, and a discussion of the findings will be presented in Section 6. The study ends by giving a conclusion of the study in terms of summarizing the major findings of the study and proposing future research directions in Section 7.

2 Methodology

The article focuses on the performance of various machine learning architectures to estimate manufacturing quality scores. The overall methodology is the data description, preprocessing, model selection, validation, and evaluation illustrated in Fig. 1.

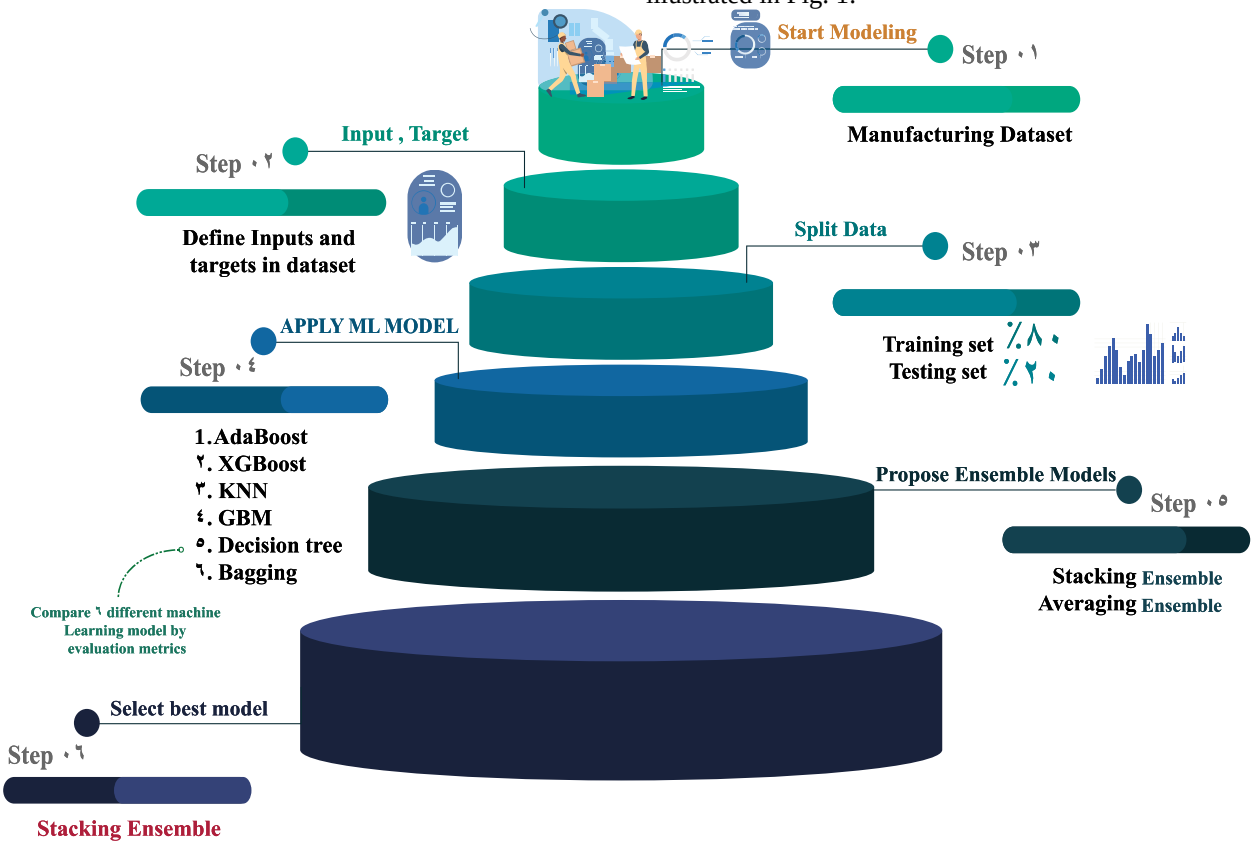


Figure 1: The overall modeling procedure for quality prediction in manufacturing.

The simulated dataset has 3,957 samples, which is a manufacturing process. Five numerical characteristics of each sample include Temperature (T), Pressure (P), the interaction of these characteristics (TxP), Material Fusion Metric (MFM), and Material Transformation Metric (MTM), and a target variable, the Quality Rating. There were no missing values, and imputation was not necessary. In order to assess this, the data were randomly split into a training set (80%, 3,165 samples) and a hold-out test set (20%, 792 samples). All features were standardized with the aid of StandardScaler of scikit-learn, and the model was fitted only to the training data to avoid information leakage.

The selection of the models and the optimization of the hyperparameters were performed with a powerful 5-fold cross-validation scheme (K-Fold (n-splits = 5, shuffle = True, random-state = 42)) on training data. This procedure offers an effective model performance estimate by training and assessing every model using five data subsets. The average RMSE and R^2 of these folds are reported as the final model performance. Wilcoxon signed-rank tests of the fold-level RMSE were conducted to statistically confirm differences in performance of the best-performing models. The model with the highest score in cross-validation was then retrained on the whole

training set and tested on the held-out test set to verify its generalization ability.

Comparisons were made between 8 ML models: six base models (AdaBoost, Bagging, Decision Tree, Gradient Boosting, K-Nearest Neighbors and XGBoost) and two ensemble models (Averaging and Stacking).. All the experiments were carried out in Python 3.10.11, using major libraries, such as pandas, numpy, scikit-learn, XGBoost, and SHAP to interpret models.

The independent variables were as follows:

- Temperature (°C) & Pressure (kPa): The fundamental process parameters.
- Temperature x Pressure (TxP): The interaction term, which represents the joint action of T and P.
- Material Fusion Metric (MFM): This is a derived measure ($T^2 + P^3$), which measures the fusion of material. Material.
- Transformation Metric (MTM): This is a derived measure ($T^3 – P^2$), which characterizes the process of material transformation.

The dependent variable, Quality Rating, serves as the comprehensive score for the final product's quality, and is a continuous, unbounded numerical score that represents a comprehensive assessment of the final product's quality, which the models are trained to predict.

Further, to evaluate the performance of each model, several evaluation criterion values were computed, such as MSE, RMSE, MAE, R2, NMSE, MDAPE, STD and VAF. MDAPE gives the median percentage error, robust to outliers, unlike standard MAE. VAF presents the percentage of variance explained, which is more easily interpreted than R^2 alone. Standard metrics (MSE, RMSE, MAE, R^2) measure overall fit, while MDAPE and VAF provide complementary, interpretable insights. Ultimately, the effectiveness of the suggested ML models was identified by contrasting and comparing the outcomes of these assessed criteria. Their equations are shown in Fig. 2.

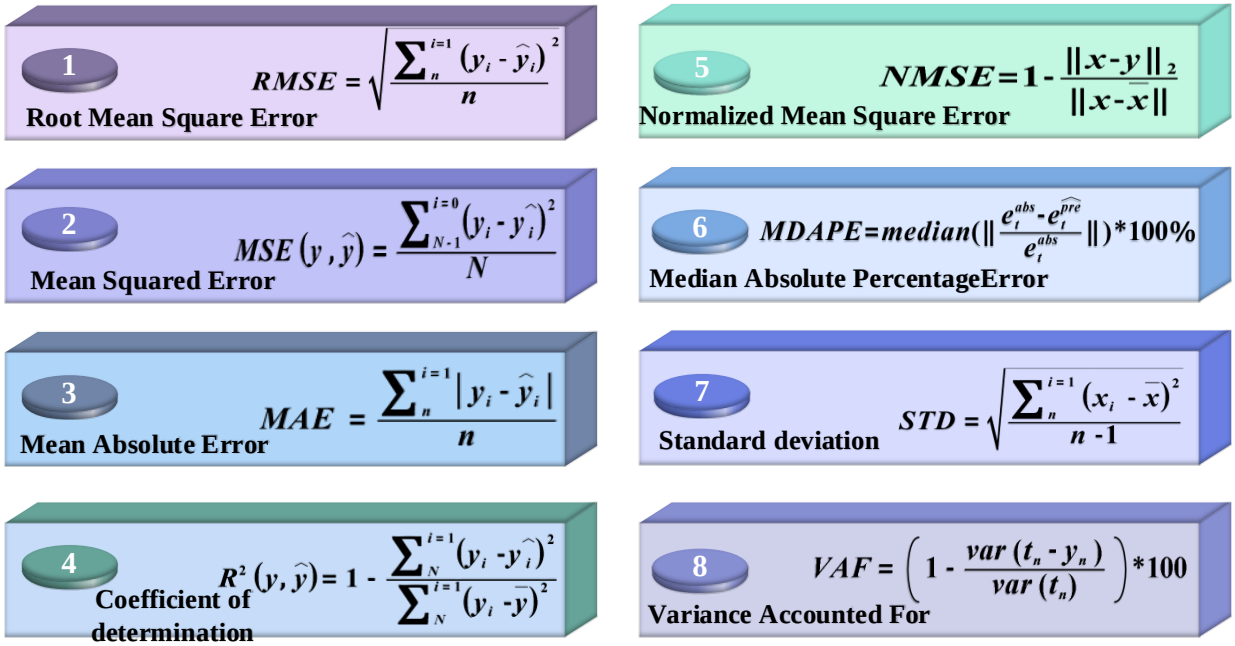


Figure 2: Equations for the evaluation of statistical metrics criteria

By accurately predicting manufacturing quality ratings, stakeholders and manufacturers can make informed decisions to optimize quality and prevent wasting costs and time. This project, which aims to use ML techniques to improve quality ratings in manufacturing, provides invaluable insights into the factors affecting manufacturing quality and offers optimization strategies for the production of manufacturing items.

3 ML models

3.1 Adaptive Boosting (AdaBoost)

AdaBoost is a sequential boosting, a type of algorithm that generates a powerful regressor by assembling a series of weak learners (e.g., decision stumps) [10]. It operates by training models repeatedly, and each time it increases the weight of false identifications [11]. This compels the following models to concentrate on instances that are more difficult to predict. The last prediction assumes that the so-called weak learners will additively make predictions, giving it high flexibility to reduce regression errors [12].

3.2 Bagging

The Bagging (Bootstrap Aggregating) is a parallel ensemble method, which tries to reduce variance. It creates numerous bootstrap copies of the initial information and educates a base model on each subset. Considering regression, the most probable prognosis will be a combination of all the individual prognoses. It works

by combining the output of individual models, which improves the stability of the predicting model by producing a more robust one that is generalized [13].

3.3 Decision Tree

A Decision Tree is used to model the predictions in a tree-like form of decisions. It divides the data into branches according to feature conditions which begin at a root node and culminate with leaf nodes that represent the ultimate predicted values. Although it is simple and interpretable, the complexity of a DT can increase with the amount of data. A tree that has the maximum possible number of branches will be able to reach high discrimination but may overfit in the absence of constraints [14].

3.4 Gradient Boosting (GB)

Gradient Boosting is a stage-wise ensemble algorithm, which builds models sequentially. It starts with a base prediction and successively includes new models that predict the negative gradient (residual errors) of the loss function of the existing ensemble. Every weak learner is tuned to correct all the mistakes of its predecessors, and its contribution is weighted by a learning rate [15]. The loss is minimized gradually until a very accurate predictor is obtained.

3.5 K-Nearest Neighbors (KNN)

KNN is an instance-based learning algorithm. In the case of regression, a target value is predicted at a new point in the data set by making the mean of the target values of the k nearest neighbors of that point in the feature space [16]. Euclidean distance is normally used in the measurement of distance. Although it is intuitively easy and efficient to represent the local patterns, its computational cost rises with the size of the dataset, because it needs to store the entire training set.

3.6 Extreme Gradient Boosting (XGBoost)

XGBoost is a regularized gradient boosting. It builds models sequentially, whereby the result of a sequence of successive trees is refined by its predecessors [17]. Some of the most significant advantages include its computational efficiency, its capability of handling missing values and the availability of both L1 and L2 regularization to control overfitting. It is an efficient and highly-utilized algorithm because it reduces a pre-defined loss function through a gradient descent optimization [18].

3.7 Averaging ensemble model

The Averaging Ensemble is a simple method that takes the mean or weighted average of predictions of several base models. This approach capitalizes on the wisdom of the crowd, in which the flaws of the individual models are likely to balance each other. It works best when there is diversity in the base models and they are accurate, which results in a reduced overall prediction variance and better performance as compared to a single model [19].

3.8 Stacking ensemble

Stacking Ensemble refers to a type of learning that combines several models to produce an improved final predictive performance [12]. The stacking model employs a meta-learner to generalize the predictions from the base models. The choice of this meta-learner is flexible and is not predefined by the method itself. The necessity of applying stacking is when multiple ML methods reveal various advantages for a certain task. In this case, the stacking ensemble method employs a discrete ML technique for specifying the efficient application of various algorithms [20]. In the Stacking model, presented in Fig. 3**,** training data are divided into different subsets, then blended and processed to produce distinctive predictions separately. Each prediction results are gathered and combined as the ultimate result of the processed training data in the study.

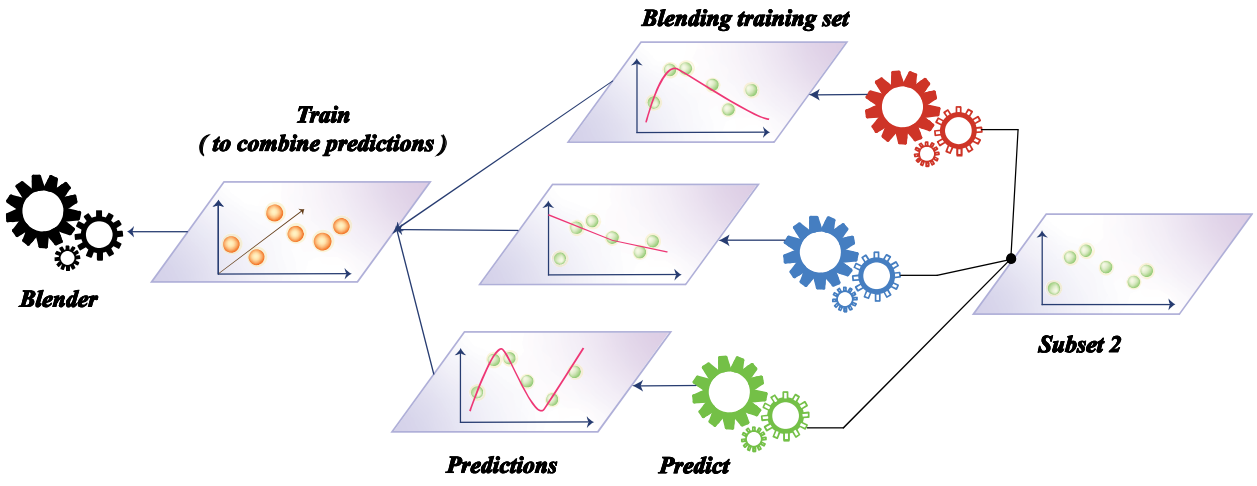


Figure 3: The steps of the Stacking Ensemble model applied for manufacturing quality rating predictions

4 Hyperparameter optimization

After the first model descriptions, hyperparameter tuning was done strictly to optimize the predictive ability of each algorithm. Optimization was done using the Grid Search technique with 5-fold cross-validation on the training data, where minimizing the Root Mean Square Error (RMSE) was the goal.

Table 2 shows the final, optimized hyperparameters of each model. This was essential to allow a fair comparison because it enabled every model to operate at its optimal level, and not in default settings.

Table 2: Hyperparameter optimization of each ML model.

Model	Hyperparameters
AdaBoost Regressor	n-estimators=150, learning-rate=0.1, random-state=42
Bagging Regressor	n-estimators=150, max-samples=0.8, max-features=0.8, random-state=42
Decision Tree Regressor	max-depth=10, min-samples-split=5, min-samples-leaf=2, random-state=42
Gradient Boosting Regressor	n-estimators=300, learning-rate=0.05, max-depth=3, subsample=0.9, random-state=42
XGBoost Regressor	n-estimators=300, learning-rate=0.05, max-depth=4, subsample=0.8, colsample-bytree=0.8, reg-lambda=1.0, objective='reg:squarederror', random-state=42
KNN Regressor	n-neighbors=7, weights='distance', metric='minkowski', p=2
Averaging Ensemble	Mean of predictions from: GB, XGBoost Regressor, KNN (same settings as above)
Stacking Regressor	Base learners: AdaBoost, Bagging, Decision Tree, GBM, XGB, KNN (all with above hyperparameters) Meta-learner: GB (n-estimators=200, learning-rate=0.05, max-depth=3, random-state=42)

These tuned models were then tested through 5-fold cross-validation, where the findings are summarized in Table 3. The cross-validation measures give a strong measure of the model's generalizability before the model testing on the final hold-out set.

Table 3: Results of 5-fold cross-validation

Model	CV-RMSE-Mean	CV-RMSE-Std	CV-R2-Mean	CV-R2-Std
AdaBoost	0.740076	0.063869	0.996689	0.000671
Bagging	0.086855	0.013201	0.999954	1.31E-05
Decision Tree	0.149076	0.014953	0.999865	3.09E-05
GBM	0.084631	0.015778	0.999956	1.41E-05
XGBoost	0.549707	0.05415	0.998187	0.000312
KNN	1.81702	0.336062	0.980178	0.005059
Stacking	0.129288	0.013833	0.9999	1.52E-05
Averaging	0.660486	0.126223	0.997362	0.000739

5 Results

This paper has compared the eight machine learning models that forecast the manufacturing quality ratings with respect to five process features. A k-fold cross-validation and a held-out test set were used to strictly compare the models and determine their performance based on a variety of statistical values (R^2, MSE, RMSE, MAE, VAF). In order to depict how closely related these variables were, a Pearson correlation heatmap was sketched as an effective color-coded visual matrix (see Fig. 4). Based on the correlation heatmap, several significant technical findings can be drawn. The obtained features, Material Fusion Metric (MFM) and Material Transformation Metric (MTM), have a near-perfect positive correlation with Temperature (0.97), which may indicate that they are functionally dependent upon it and may not be necessary to model. In addition, MFM and MTM are strongly correlated with each other (0.98), which means high levels of multicollinearity that may render some model interpretations unstable. Importantly, the target variable, Quality Rating, has moderate to strong negative correlations with Temperature (-0.46) and MFM (-0.51), and even stronger with MTM (-0.58), which are, in its turn, confirmed by the SHAP analysis.

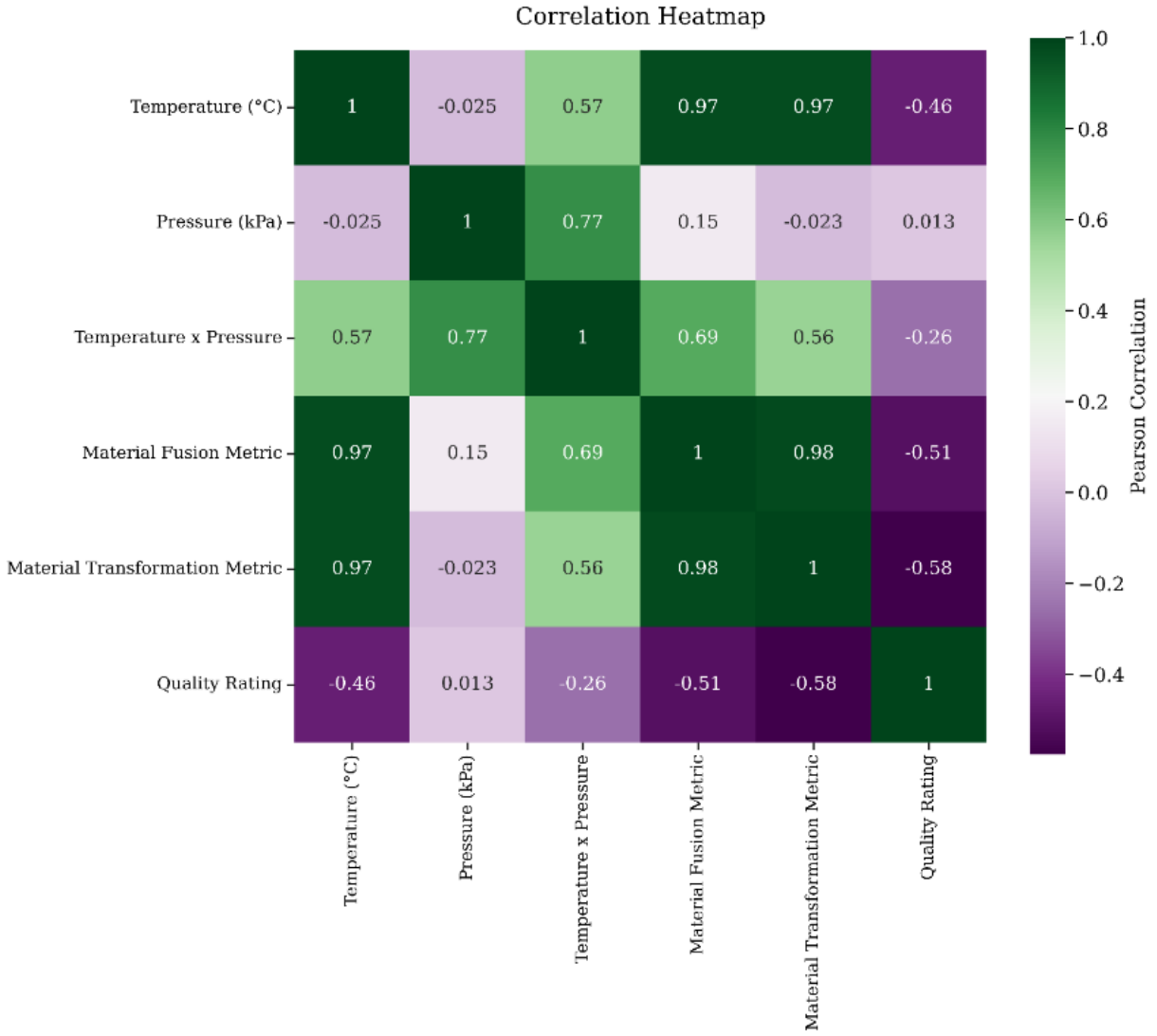


Figure 4: Pearson Correlation Heatmap for examining the relationship between the independent variables

Preliminary analysis of the six base models (AdaBoost, Bagging, DT, GB, KNN, XGBoost) showed that the majority of them demonstrated high and comparable results on the test sample, with R^2 and VAF metrics staying nearly at 0.999. Nevertheless, a closer examination of error measures showed that KNN is the most precise single model, with the lowest test errors (MSE: 0.006, RMSE: 0.078, MAE: 0.018). Conversely, AdaBoost and XGBoost showed significantly higher rates of error, which makes them the weakest among the base models (Table 4).

Table 4: Error metrics obtained by using the primary proposed models.

Model	Dataset	MSE	RMSE	MAE	R^2	NMSE	MDAPE	STD_dev	VAF
AdaBoost	Train	0.499	0.706	0.42	0.997	0.003	0.26	13.066	0.997
	Test	0.624	0.79	0.458	0.996	0.004	0.26	12.774	0.996
Bagging	Train	0.003	0.053	0.01	0.999	1.62E-05	3.34E-07	13.075	0.999
	Test	0.011	0.105	0.024	0.999	6.94E-05	5.14E-07	12.622	0.999
DT	Train	0	6.63E-07	0	0.999	2.57E-15	1.39E-07	13.085	0.999
	Test	0.012	0.107	0.022	0.999	7.26E-05	7.04E-07	12.612	0.999
GB	Train	0.002	0.047	0.015	0.999	1.31E-05	0.001	13.083	0.999
	Test	0.012	0.108	0.032	0.999	7.27E-05	0.001	12.602	0.999
XGB	Train	0.001	0.036	0.01	0.999	7.55E-06	0	13.085	0.999
	Test	0.461	0.679	0.147	0.997	0.003	0	12.695	0.997
KNN	Train	0.005	0.069	0.014	0.999	2.75E-05	1.17E-08	13.079	0.999
	Test	0.006	0.078	0.018	0.999	3.86E-05	2.09E-08	12.62	0.999

Considering Fig. 5 below, a comparison of data values obtained through the integrative ML parameters was conducted. This plot depicts the initial suggested models, which include AdaBoost, Bagging, DT, GB, KNN and XGBoost and distances of each to the target value. Each model trained on 80% and tested on 20% of the data. The findings of this plot indicate that the values of all proposed models overlap with the actual value, with KNN and Bagging showing the closest alignment.

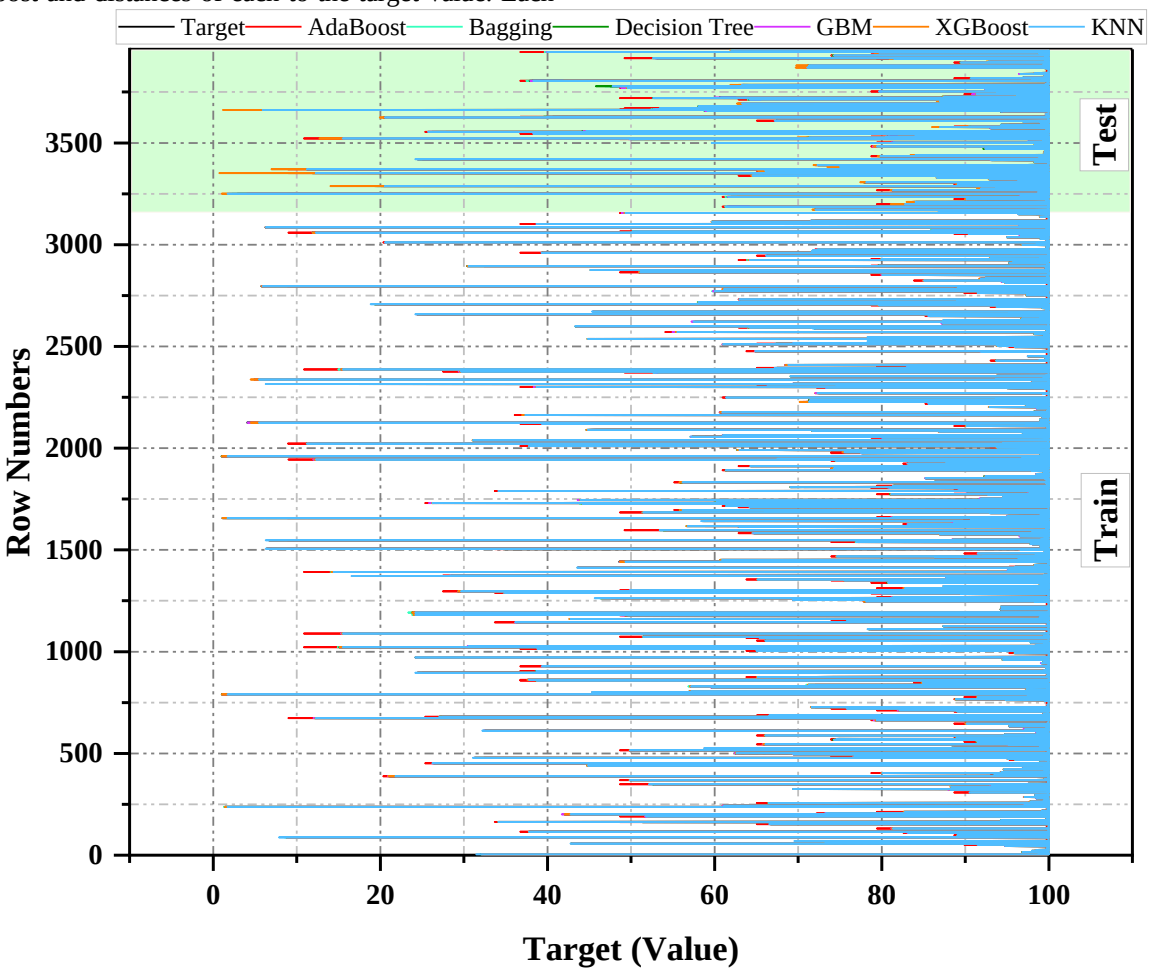


Figure 5: Target value plot comparing the primary proposed ML models for manufacturing quality predictions.

Additionally, as shown in Fig. 6, the error values of each proposed ML model are illustrated for both train and test data. A model with the least error values (i.e., closest to zero) is considered the best predictor among the employed ML models. This scheme visually evaluates the accuracy of model predictions. Based on this, among the base models, KNN emerged as the best model, exhibiting the lowest error values, while GB ranked second with the next lowest error. Conversely, AdaBoost was the least accurate, with the highest error values. This finding aligns with the subsequent histogram plot (Fig. 7), where the error density is analyzed. As evident in this figure, KNN and GB exhibit the lowest error density, making them the superior models, while AdaBoost shows a notably high error density, indicating its weaker predictive performance.

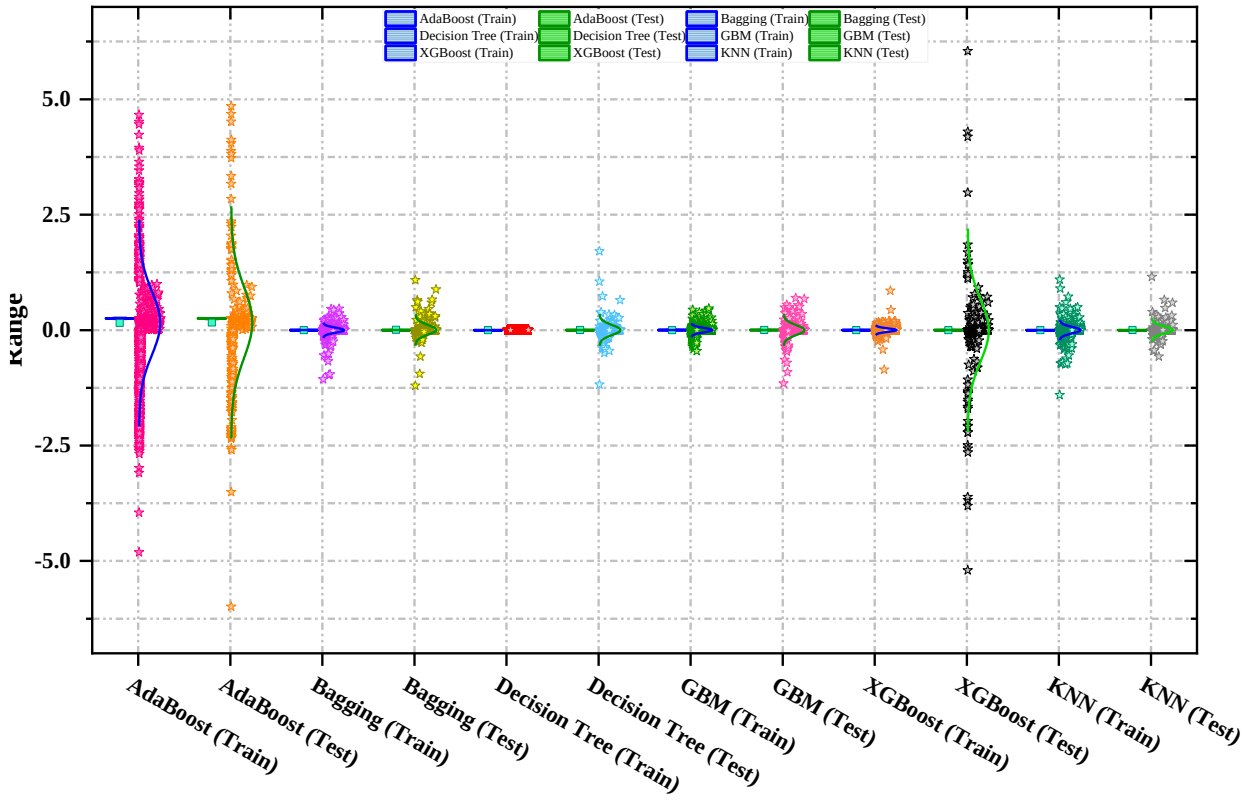


Figure 6: Boxplot of the error values of the proposed ML models for both train and test data in manufacturing quality prediction.

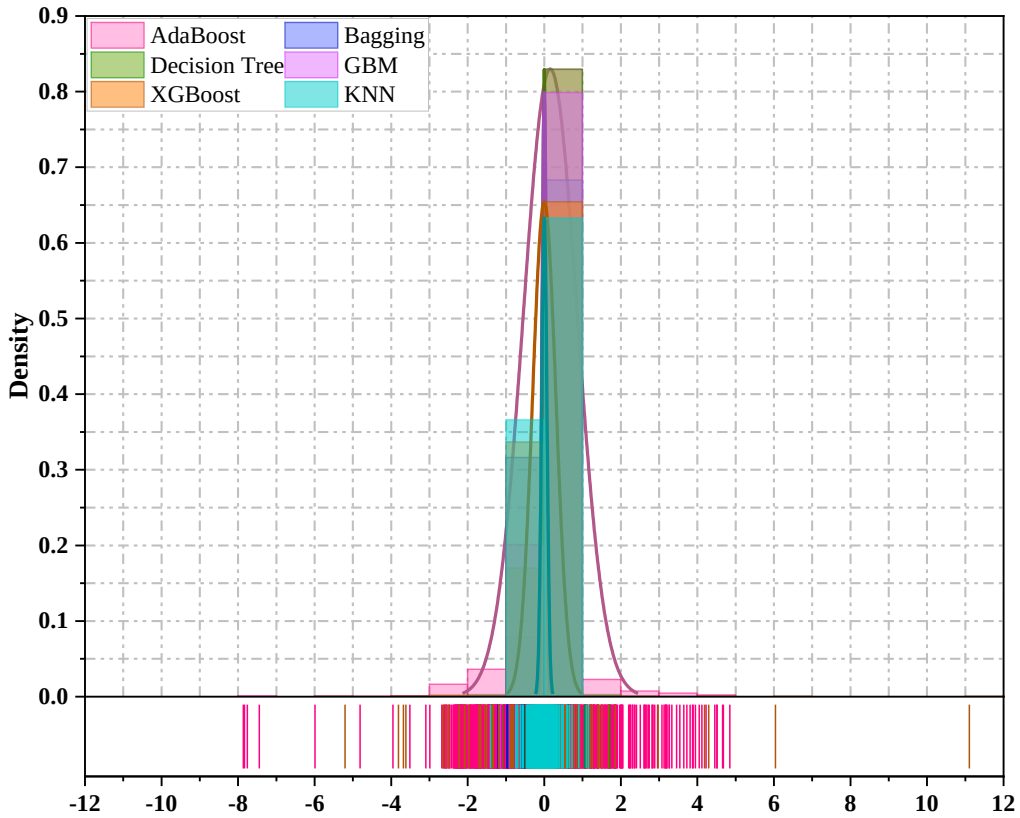


Figure 7: Histogram plot of the errors of the proposed ML models in manufacturing quality prediction

In addition to the output of Table 4, another graphic comparison has been illustrated based on eight significant error metrics below (see Figs. 8 and 9), fulfilling a precise and visual analysis of the primary models' performance in the present work. Clear from these plots, the histograms of each model's outcome are almost consistent with the

abovementioned results. A quick look into RMSE, MAE, MSE, and NMSE plots showed that KNN, GB, DT, and Bagging had almost similar performance, excluding AdaBoost and XGBoost, in the testing data. In Fig. 9, with respect to R^2, STD, VAF, and MDAPE histograms, all the models performed roughly the same once more.

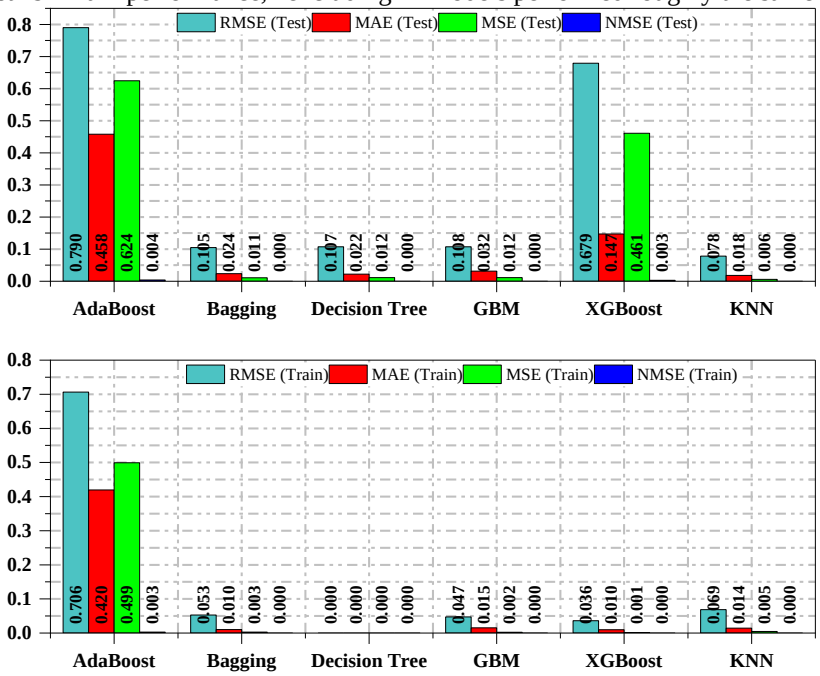


Figure 8: RMSE, MAE, MSE, NMSE histograms based on the primary proposed models' performance for both train and test data.

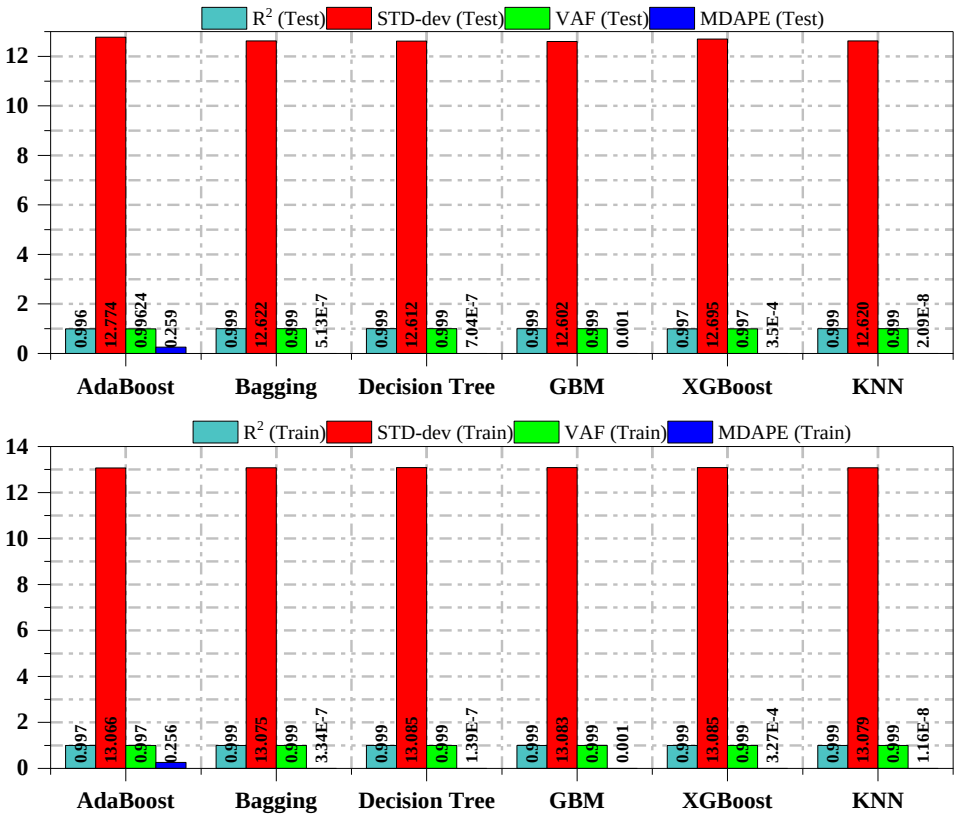


Figure 9: R^2, STD, VAF, MDAPE histograms based on the primary proposed models' performance for both train and test data

Since the performance of the top base models was closely matched, two sophisticated ensemble methods, Averaging and Stacking, were applied to determine whether the predictive accuracy could be improved. The Stacking ensemble that employed the GB meta-learner was superior. It had the lowest error measures on the test

set (MSE: 0.004, RMSE: 0.065, MAE: 0.016) with a perfect R^2 and VAF score of 0.999 (Table 5). This proves the hypothesis that a stacking ensemble can utilize the strengths of many base learners to produce the strongest and most accurate predictions.

Table 5: Error metrics of the two additive ensemble models.

Model	Dataset	MSE	RMSE	MAE	R^2	NMSE	MDAPE	STD-dev	VAF
Averaging	Train	0.017	0.131	0.078	0.999	0	0.047	13.083	0.999
	Test	0.033	0.181	0.092	0.999	0	0.047	12.656	0.999
Stacking	Train	0.002	0.048	0.009	0.999	1.37E-05	6.90E-07	13.089	0.999
	Test	0.004	0.065	0.016	0.999	2.69E-05	8.92E-07	12.628	0.999

Fig. 10 demonstrates the error distribution histograms of both Averaging and Stacking ensemble models, and gives a clear visual justification that the predictive performance of the Stacking ensemble model is superior to that of the Averaging ensemble model. The Stacking ensemble has a significantly smaller error profile centered around zero, meaning that it consistently predicts with a minimal error. Conversely, the Averaging model shows more spread of errors, which expresses greater variation in prediction. This specific difference highlights the effectiveness of the Stacking meta-learner in integrating the capabilities of its base models not only to reduce average errors but also to obtain a higher degree of reliability in predicting the quality of manufacturing.

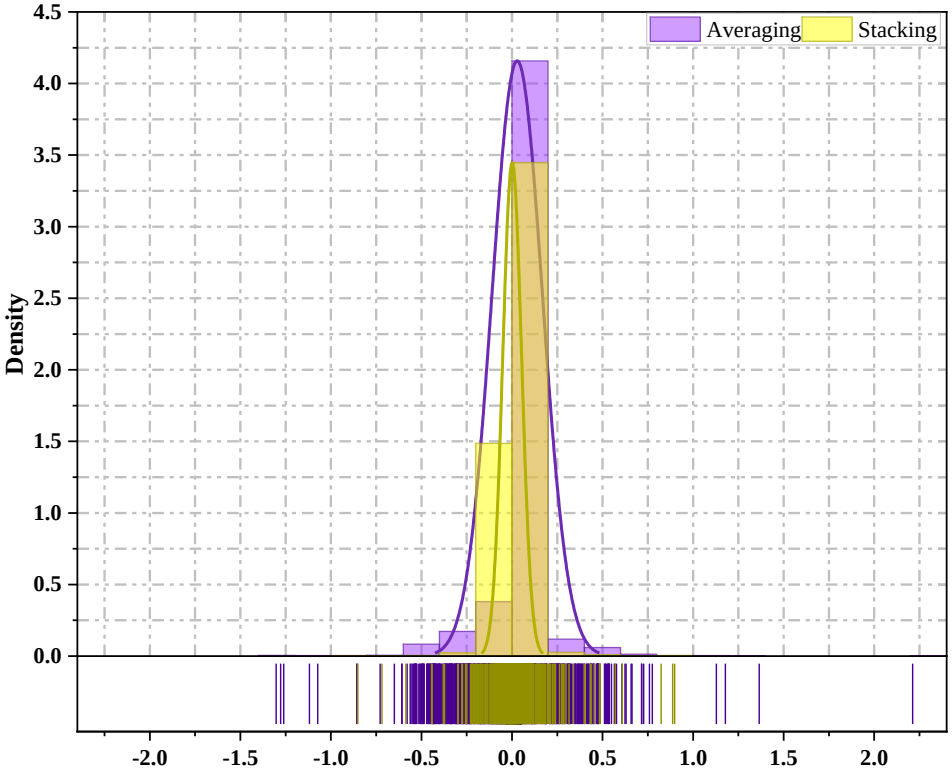


Figure 10: Histogram plot of the errors of the two additive ensemble models in manufacturing quality prediction.

For more scrutiny, another comparison for the accuracy and efficiency of the applied models in this study was carried out through examining R^2 scatter plots of these models. R^2 results as a significant error measure are presented in Fig. 11 and demonstrate the extent to which the predictions of the employed ML models correspond to the actual data. As seen in this figure, a model is deemed superior and more accurate if its prediction values nearly align with the norm line (when $R^2 = 1$). As shown, all the models served this purpose; i.e., their forecasts closely matched the norm line, excluding AdaBoost and XGBoost, which were weaker models. Hence, higher R^2 values were obtained in KNN, Stacking, Bagging, DT, GB, and Averaging, all of whose R^2 values equaled approximately 0.999 for both the train data and test data. However, the lowest RMSE was merely for the Stacking model, meriting being the best fit model; i.e., 0.048 and 0.065, respectively, for train and test data. As a result, staking outperformed the other models. This was also in accordance with the results in the above-mentioned training and testing tables' outcomes.

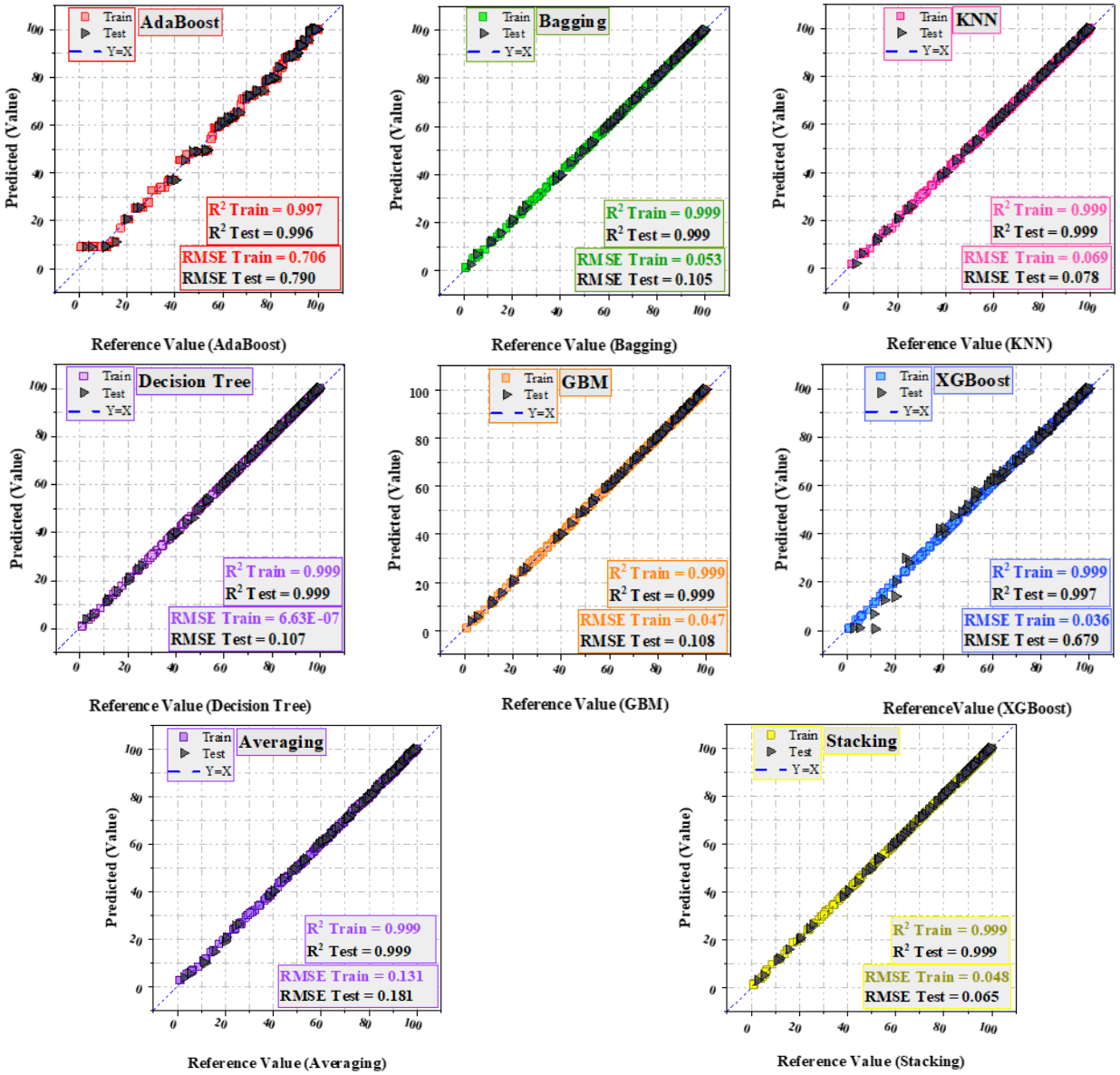
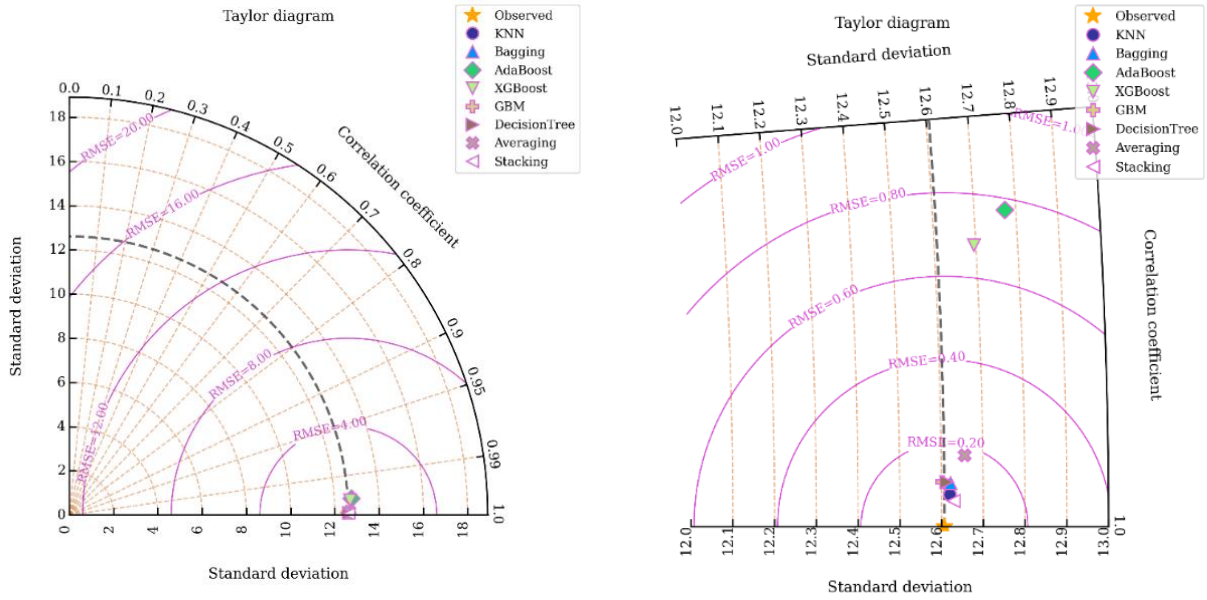


Figure 11: Comparing the coefficient of determination (R^2) for the individual proposed ML models

Another graphical diagram, named the Taylor diagram, has been used here for comparing the recommended models' performance. This chart compares models against one another based on the model accuracy using correlation coefficient, STD and RMSE [21]. The performance of the models is presented as a circle on this diagram where the better performance of the models is determined by the nearest points to the reference point of the reference [22]. The Taylor diagram is depicted in Fig. 12 for predicting manufacturing quality. According to the diagram, the relative performance of the models can be succinctly assessed using the distance between the models and the reference point of the models, which is an ideal model. The best-performing models, including Stacking, Bagging, GBM, and Decision Tree, are those that are closest to the reference point, with a high correlation coefficient (>0.99), a standard deviation that is nearly equal to that of the observations, and thus the lowest values of RMSE. Conversely, AdaBoost and XGBoost are further away, which reflects the fact that they have a high RMSE and weak correlation, which is consistent with their designation as the poorer performers in the study. The Stacking ensemble consistently appears within this top-performing cluster, providing a visual confirmation of its optimal balance between high accuracy and low error, as substantiated by the quantitative results.



Test Test (Zoom)

Figure 12: Taylor diagrams for the models' comparison based on RMSE, STD, R metrics

A series of pairwise Wilcoxon signed-rank tests was performed to statistically justify the performance differences in the cross-validation folds between the best-performing Stacking model and all other models. Table 6 shows that no pairwise comparison showed a statistically significant difference at the 0.05 level of significance (all p-values were larger than 0.05). This insignificance can probably be explained by the fact that the models already demonstrate extremely high and comparable performance (R^2 = 0.999), and it is hard to statistically prove that the difference is really present in this test. Thus, although the Stacking model is arguably the most accurate predictor based on its lower error scores, the Wilcoxon test indicates that its performance, in the strict statistical sense, does not differ significantly from the other base models, such as KNN and GBM, in this particular experimental setting.

Table 6: Results of the pairwise Wilcoxon tests.

Reference model	Compared model	Wilcoxon statistic	p-value	Significant-at-0.05
Stacking	AdaBoost	0	0.0625	FALSE
Stacking	Bagging	0	0.0625	FALSE
Stacking	Decision Tree	1	0.125	FALSE
Stacking	GBM	0	0.0625	FALSE
Stacking	XGBoost	0	0.0625	FALSE
Stacking	KNN	0	0.0625	FALSE
Stacking	Averaging	0	0.0625	FALSE

The Stacking model, as the best performing model, was examined with SHAP to explain its predictions and measure the effect of features to make sure that it was useful in practice. The analysis showed a definite order of feature value (Fig. 13). The most prevailing one was the Material Transformation Metric (MTM), then Temperature (T). The impacts of the Material Fusion Metric (MFM), the interaction term (TxP), and Pressure (P) were much lower.

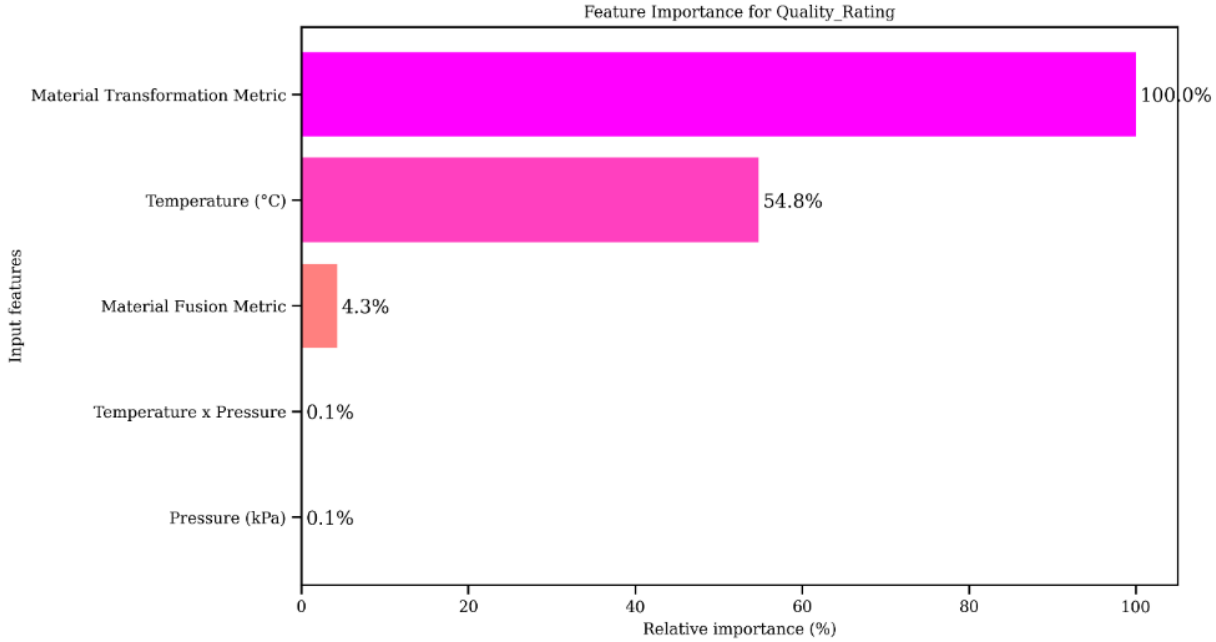


Figure 13: Features' importance for manufacturing quality rating predicted by the Stacking model

These relationships were further clarified by the SHAP summary plot (Fig. 14) [23], which revealed that greater values of MTM and T were consistently associated with higher predicted quality ratings. The individual SHAP dependence curves (Fig. 15) confirmed the steady and monotonic relationship of MTM and T with the output as opposed to the more dispersed and non-influential trends of MFM, TxP, and P.

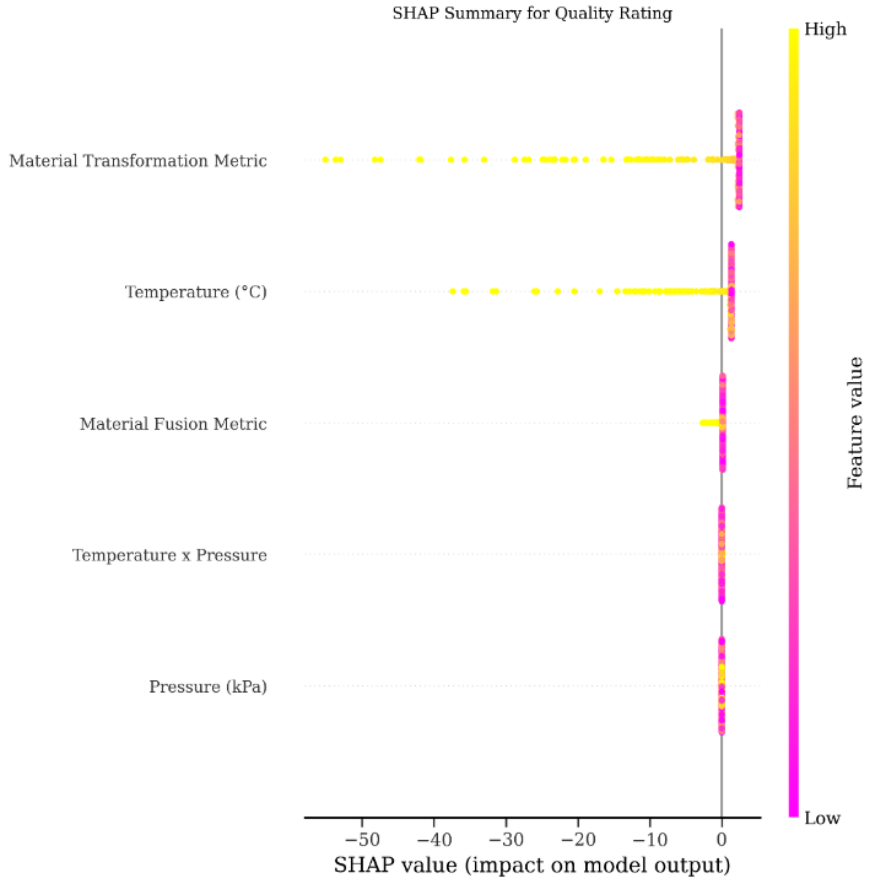


Figure 14: SHAP values providing an overview of the impacts of each manufacturing feature on the quality rating predictions by the selected Stacking model

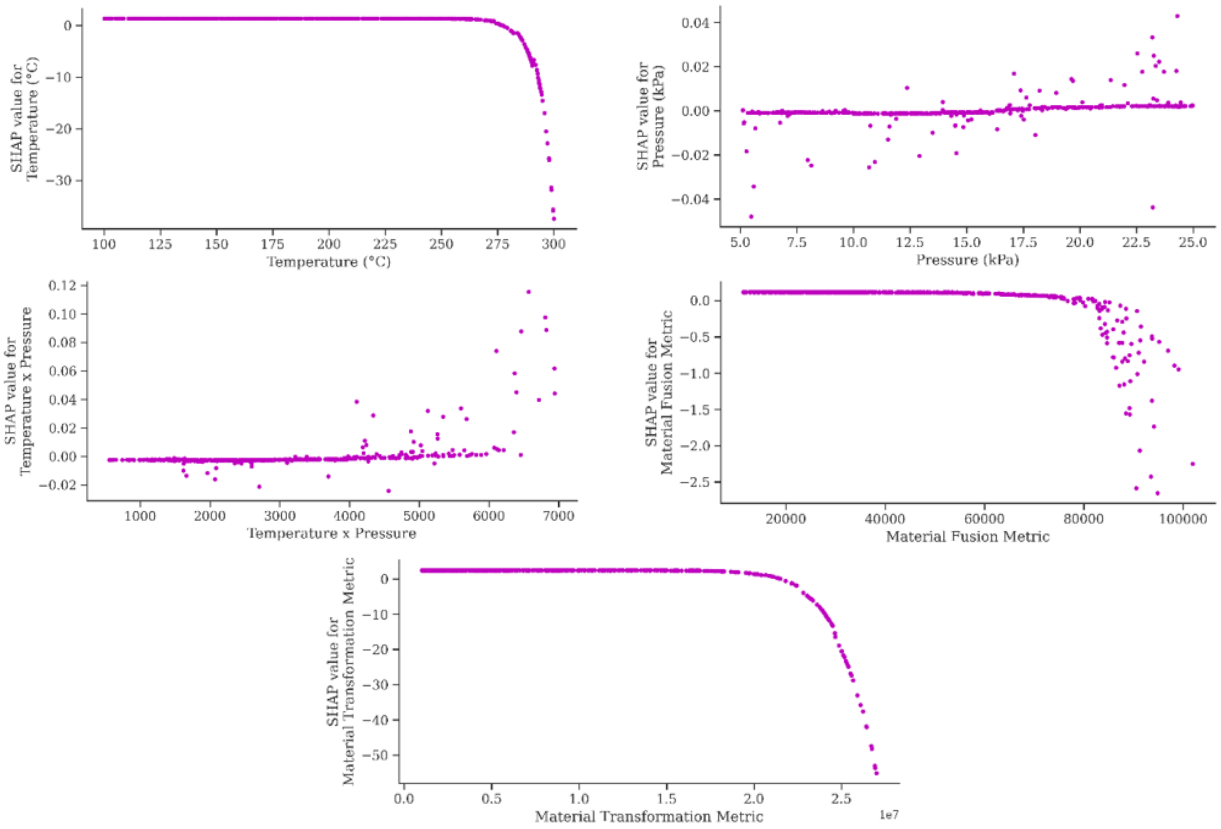


Figure 15: SHAP Subplots by Stacking model.

6 Discussion

This paper shows that the Stacking Ensemble model is the most effective for predicting manufacturing quality ratings, and the model provides almost perfect performance (R^2: 0.999) and low errors (RMSE: 0.065). This observation confirms the paper's hypothesis that a meta-learner can successfully combine the power of different base models to produce a more powerful predictor.

Compared to the existing literature (Table 1), this work's results contribute to the field in two important aspects. First, although earlier research, such as Li et al. [1] and Msakni et al. [6], achieved success in applying single or combined models to their particular situations, it finds that a general Stacking ensemble offers a final margin of performance improvement that could be pivotal to high-stakes, zero-defect manufacturing objectives. Second, it discusses the importance of model interpretability as emphasized by Link et al. [3] and Weichert et al. [4]. Using SHAP analysis on the best model, this study goes beyond a black box to numerically demonstrate that the most influential features are Material Transformation Metric (MTM) and Temperature (T), which can guide manufacturers to control their processes.

The Wilcoxon test revealed no statistically significant difference between the base models and the Stacking ensemble, which is probably because their performance is already exceptionally high. However, the Stacking model demonstrates a consistent tendency of having the lowest error metrics, which is why it can be selected as the most reliable one when it comes to the practical applications where marginal gains are of the essence. Therefore, the paper not only proposes an excellent predictive model but also offers a framework for exploiting the collective model intelligence with explainable outputs for decision-making in industry.

7 Conclusion

This research was able to create a high-precision predictive model of manufacturing quality ratings. After a thorough analysis of eight machine learning models, the Stacking Ensemble was determined to be the best model with almost perfect prediction (R^2 = 0.999) and minimal error rates (MSE = 0.004, RMSE = 0.065). The combination of several base learners via a meta-learner turned out to be the most effective strategy. Moreover, SHAP analysis offered essential interpretability, as MTM and Temperature were found to be the most significant features of quality, which could be used to introduce practical recommendations in terms of optimizing the process.

Nevertheless, this work has some limitations despite its high performance. Simulated data though controlled might not accurately reflect the noise and complexity of real-world production environments. On the other hand, the model was trained on a particular number of five features, which may not capture other contributory variables in real-life manufacturing contexts.

Further studies should concentrate on testing such results with actual industrial data to complete the practical generalizability of the model. It is also suggested to widen the feature set with more dynamic process parameters and

external factors. Future research may consider more sophisticated deep learning designs and online learning to enhance models in adaptive manufacturing systems. Finally, the adoption of more explainable AI (XAI) techniques can bring additional and more transparent insights into the cause-and-effect patterns in the production process.

Acknowledgements

This work was supported by.

Project 1: YB202215 Key Scientific Research Project of Jingchu University of Technology

Project 2: Provincial Teaching and Research Project for Universities in Hubei Province (2022450)

Project3: School-level Research Platform of Jingchu University of Technology:Data Analysis Science Laboratory

Project4: HX20220171 Horizontal Scientific Research Project of Jingchu University of Technology

Project5: Research project of Jingchu University of Technology (YB202301, JX2022-007)

Project6: Education Science Planning Project of Jingmen City (JMG2022004)

Project7: China Association for Educational Technology Project (XJJ202205022）

Project8: School Level Project, QN202419

References

[1] X. Li, Z. Huang, and W. Ning, “Intelligent manufacturing quality prediction model and evaluation system based on big data machine learning,” *Computers and Electrical Engineering*, vol. 111, p. 108904, 2023. https://doi.org/10.1016/j.compeleceng.2023.108904

[2] S. Sankhye and G. Hu, “Machine learning methods for quality prediction in production,” *Logistics*, vol. 4, no. 4, p. 35, 2020. https://doi.org/10.3390/logistics4040035

[3] P. Link *et al.*, “Capturing and incorporating expert knowledge into machine learning models for quality prediction in manufacturing,” *J Intell Manuf*, vol. 33, no. 7, pp. 2129–2142, 2022. https://doi.org/10.1007/s10845-022-01975-4

[4] D. Weichert, P. Link, A. Stoll, S. Rüping, S. Ihlenfeldt, and S. Wrobel, “A review of machine learning for the optimization of production processes,” *The International Journal of Advanced Manufacturing Technology*, vol. 104, no. 5, pp. 1889–1902, 2019. https://doi.org/10.1007/s00170-019-03988-5

[5] F. Psarommatis and V. Azamfirei, “Zero Defect Manufacturing: A complete guide for advanced and sustainable quality management,” *J Manuf Syst*, vol. 77, pp. 764–779, 2024. https://doi.org/10.1016/j.jmsy.2024.10.022

[6] M. K. Msakni, A. Risan, and P. Schütz, “Using machine learning prediction models for quality control: a case study from the automotive industry,” *Computational management science*, vol. 20, no. 1, p. 14, 2023. https://doi.org/10.1007/s10287-023-00448-0

[7] B. Yang, J. Huang, and Y. Chen, “The relationship between ESG ratings and digital technological innovation in manufacturing: Insights via dual machine learning models,” *Financ Res Lett*, vol. 71, p. 106362, 2025. https://doi.org/10.1016/j.frl.2024.106362

[8] P. Akbari, M. Zamani, and A. Mostafaei, “Machine learning predictions of spatter behavior in LPBF additive manufacturing,” *Materialia (Oxf)*, vol. 38, p. 102268, 2024. https://doi.org/10.1016/j.mtla.2024.102268

[9] V. Azamfirei, F. Psarommatis, and Y. Lagrosen, “Application of automation for in-line quality inspection, a zero-defect manufacturing approach,” *J Manuf Syst*, vol. 67, pp. 1–22, 2023. https://doi.org/10.1016/j.jmsy.2022.12.010

[10] S. Leiprecht, F. Behrens, T. Faber, and M. Finkenrath, “A comprehensive thermal load forecasting analysis based on machine learning algorithms,” *Energy Reports*, vol. 7, pp. 319–326, 2021. https://doi.org/10.1016/j.egyr.2021.08.140

[11] C. Ying, M. Qi-Guang, L. Jia-Chen, and G. Lin, “Advance and prospects of AdaBoost algorithm,” *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, 2013. https://doi.org/10.1016/S1874-1029(13)60052-X

[12] H. Peng, J. Xiong, C. Pi, X. Zhou, and Z. Wu, “A dynamic multi-objective optimization evolutionary algorithm with adaptive boosting,” *Swarm Evol Comput*, vol. 89, p. 101621, 2024. https://doi.org/10.1016/S1874-1029(13)60052-X

[13] A. Almomani *et al.*, “Age and Gender Classification Using Backpropagation and Bagging Algorithms.,” *Computers, Materials & Continua*, vol. 74, no. 2, 2023. DOI: 10.32604/cmc.2023.030567

[14] U. R. Konduru, A. P. Nagarajan, and C. V. S. Sai, “An improved performance of reversible data hiding in encrypted images using decision tree algorithm,” *Eng Appl Artif Intell*, vol. 137, p. 109100, 2024. https://doi.org/10.1016/j.engappai.2024.109100

[15] R. Huang, C. McMahan, B. Herrin, A. McLain, B. Cai, and S. Self, “Gradient boosting: A computationally efficient alternative to Markov chain Monte Carlo sampling for fitting large Bayesian spatio-temporal binomial regression models,” *Infect Dis Model*, vol. 10, no. 1, pp. 189–200, 2025. https://doi.org/10.1016/j.idm.2024.09.008

[16] L. Hardesty, “Explained: neural networks,” *MIT News*, vol. 14, 2017.

[17] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785

[18] B. Kıyak, H. F. Öztop, F. Ertam, and İ. G. Aksoy, "An intelligent approach to investigate the effects of container orientation for PCM melting based on an XGBoost regression model," *Eng Anal Bound Elem*, vol. 161, pp. 202–213, 2024. https://doi.org/10.1016/j.enganabound.2024.01.018

[19] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, Springer, 2000, pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1

[20] U. Arif, C. Zhang, S. Hussain, and A. R. Abbasi, "An efficient interpretable stacking ensemble model for lung cancer prognosis," *Comput Biol Chem*, vol. 113, p. 108248, 2024. https://doi.org/10.1016/j.compbiolchem.2024.108248

[21] K. E. Taylor, "Summarizing multiple aspects of model performance in a single diagram," *Journal of geophysical research: atmospheres*, vol. 106, no. D7, pp. 7183–7192, 2001. https://doi.org/10.1029/2000JD900719

[22] M. Ehteram, A. N. Ahmed, P. Kumar, M. Sherif, and A. El-Shafie, "Predicting freshwater production and energy consumption in a seawater greenhouse based on ensemble frameworks using optimized multi-layer perceptron," *Energy Reports*, vol. 7, pp. 6308–6326, 2021. https://doi.org/10.1016/j.egyr.2021.09.079

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv Neural Inf Process Syst*, vol. 30, 2017.

GCN-Att-GRU: Spatio-temporal Graph Convolution with Multi-Head Attention for Multi-step Urban Water Demand Forecasting

Muhua Hu
Mathematics and Digital Sciences Department of Chengyi College, Jimei University, XiaMen, FuJian,361021, China
E-mail: humuhuacy@163.com

Keywords: urban water demand prediction, china, temporal and spatial attributes, GCN-Att-GRU Model, multi-head attention model

Received: July 12, 2025

Urban water resource planning as well as the supply-demand balance are directly affected by immediate demand forecasts, making them vital for water resource management. The present approaches to water demand forecasting just consider temporal variables and do not take account of the possible impact of geographical characteristics on them. The reason for this is because short-term urban water demand forecasts are affected by several variables, many of which display complex nonlinear dynamic features. Predictions end up being inaccurate because of this. This research aims to address this problem by presenting a model that considers geographical and temporal characteristics in order to predict urban water consumption in the near term. The first step is to detect and fix anomalies using the Prophet model. In order to generate an adjacency matrix among variables, we use a maximum information coefficient, and to extract spatial attributes between variables, we use a graph convolutional neural network. Afterwards, a multi-head attention method is used to enhance crucial aspects of water consumption statistics while reducing the influence of unimportant components. The next phase involves projecting urban areas' short-term water demands using a Gated Recurrent Unit (GRUs). The suggested model achieves a Mean Absolute Percentage Error (MAPE) of 2.98% (±0.05) across a 7-day forecasting horizon, which is a significant 7.65% decrease in MAPE when compared to the optimized baseline. The mechanistic essentiality of the MIC-GCN and the MHA mechanism's integration for this enhanced performance has been confirmed by ablation experiments. Based on these findings, the GCN-Att-GRU may be confidently used to predict the operational water demand in cities.

Povzetek: Raziskava predstavi model, ki združi prostorske (grafovne) in časovne informacije za natančnejše kratkoročno napoved mestne porabe vode ter doseže bistveno nižjo napako (MAPE ~3%) kot obstoječe metode.

1 Introduction

Per capita freshwater availability is poor in China, even though the country has a lot of freshwater resources overall [1]. The rapid growth of China's urban areas has resulted in water shortages for the majority of cities. To address this issue, it is essential to carefully manage how urban residents utilize water in their homes [2]. The water consumption of urban residential areas is characterized by complicated nonlinear dynamics and is affected by a multitude of variables [3]. When it comes to complex water use trends, traditional approaches relying on human experience fall short [4]. However, urban water demand forecast may help uncover consumption patterns and guide efficient water resource allocation and use. There are three types of urban water demand predictions: short-term, medium-term, with long-term [5]. The immediacy and urgency of the short-term predictions make them especially important for water resource management [6]. Cities can increase the efficiency of their water delivery systems, save costs and energy usage, and plan for the future with more precision when they can accurately estimate water demand in the near term [7].

Several methods exist at the moment for forecasting future water demand. Grey models, SVMs, RFs, regression analysis designs, and traditional BPNNs are all part of this category [8]. These methods may be effective in certain contexts, but they lack the ability to generalize or make accurate predictions [9]. Predicting urban water demand has become more popular with the use of neural network along with deep learning techniques, thanks to recent advancements in this field. When it comes to predicting future water demands, the GRU model has performed quite well [10]. It excels at handling data with a high temporal resolution, as well as data with abrupt changes and uncertainty [11]. Additional details, such as the current day and national holidays, may

also be included. However, there are still major obstacles to overcome, such as the effects of weather and information noise on the precision of predictions [12].

Current methods still mostly focus on extracting time series data's temporal properties, ignoring the impact of spatial variables on these temporal elements, despite substantial progress. Currently, multi-site analysis is the main emphasis of research on spatiotemporal characteristics [13].

Previous work has taken into account the impact of these aspects on water consumption data over time, but current efforts primarily aim at establishing correlations between spatial qualities at different observation locations rather than between the geographical features that comprise clear variables [14].

Given the above, GCN-Att-GRU, a model for attention processes in graph convolutional neural networks that include both spatial and temporal variables, is presented by the authors of this research [15]. Despite its complexity, this approach provides more accurate short-term predictions of urban water demand than earlier techniques [16]. Finding and fixing outliers is the first step in improving water data quality using the Prophet model [17]. Next, we use the greatest information coefficient to calculate the relationship between the weather and the water use statistics [18]. The next step is to use a multi-head attention mechanism to ensure that the graph convolutional network extracts an equal amount of geographic information on water usage and weather conditions [19]. Finally, the processed data is input into a three-layer GRU with residual connections added to improve the accuracy of the water demand estimate [20].

The following is the structure of the paper: The sources of the processed urban water data are detailed in Section 2. Section 3 explains the steps taken by the GCN-Att-GRU model. In Section 4, we examine the experimental data to verify the model's correctness. Section 5 wraps up the investigation by discussing the research outcomes.

2 Related work

Drawing on the work of researchers in [21], the authors of that paper develop a model that takes geographical and temporal factors into account in order to predict urban water demand for the near future. The first step is to detect and fix anomalies using the Prophet model. In order to generate an adjacency matrix among variables, we use a maximum information coefficient, and to extract spatial attributes between variables, we use a graph convolutional neural network. Afterwards, a multi-head attention method is used to enhance crucial aspects of water consumption statistics while reducing the influence of unimportant components. Predicting urban regions' short-term water consumption using a three-layer GRU structure is the next stage. By reducing the average absolute percentage error by 1.868-2.718% compared to existing prediction models, the study suggests a hybrid model that improves forecast accuracy and effectiveness. This one may aid municipalities in making better use of their water resources and will also pave the way for further research.

In a revolutionary move, the authors of [22] use data pre-processing and an Artificial Neural Network improved using the Backtracking Search Process to forecast monthly water consumption based on previous use. Information on monthly water use in Gauteng Province, South Africa, from 2007 to 2016 was utilized to design and evaluate the approach. Data pre-processing approaches were crucial for improving data quality before creating the prediction model. The best result was achieved by the BSA-ANN model, which had an efficiency coefficient of 0.979 and a root-mean-square error of 0.0099 mega liters. When compared to the Crow Search Algorithm, it was shown to be more efficient and dependable in terms of error scale. In sum, that article introduces a novel use case for the BSA-ANN hybrid model, which demonstrates how to accurately forecast water demand in a metropolis hit hard by both climate change with population boom.

In [23], the authors introduced a probabilistic approach to forecasting that accounts for systemic uncertainty to foretell future water use trends. That approach uses statistical methods to predict the distribution of possible future water demand situations. Better scheduling and handling of water resources is achievable because to the many options presented to decision-makers by that study's results. They developed a hybrid model for hourly water demand estimation using conformal predictions after comparing several machine learning approaches. That model integrates CNNs with bidirectional long short-term memories. Furthermore, they discuss important aspects of constructing a probabilistic forecasting framework, such as picking the right data and the right parameters for the model. For probabilistic water demand forecasting, the suggested model has been tested and shown to work in the actual world. Deterministic predictions improved by 10% and probabilistic forecasts by 26.7%, according to the results. Results show that method has promise for better resource management and decision-making.

Data preprocessing has the potential to fix the outlier issue, as stated by the authors of [24]. The results demonstrate that the proposed hybrid model outperforms similar machine learning models when it comes to prediction accuracy and resilience in the face of prediction uncertainty. Feature engineering improves the data input dimension before the LightGBM (Light Gradient Boosting Machine) method is used to predict

future water demand. When it comes to missing value interpolation jobs, the findings show that cubic polynomial interpolation is better than both the Prophet model and the linear technique. The LightGBM model does a great job of predicting future patterns in water demand and has great forecasting performance overall. The test dataset has an assessment indication of 4.28% for MAPE (mean absolute percentage error) with an indicator of 0.94 for NSE (Nash-Sutcliffe efficiency coefficient). With these metrics as a foundation, water supply firms may make educated guesses about the near future.

The authors of [25] the classic techniques of water demand forecast are encountering difficulties, necessitating the development of new methods and instruments to enhance the precision and efficacy of their projections. From 1983 to 2024, the pertinent research on urban water demand projection was visually analyzed in the CNKI database with the Web of Science core literature database using CiteSpace software, bibliometrics, and visualization analysis methodologies. In order to pinpoint areas of concentrated research and developing patterns in urban water demand prediction, they will be using visual tools like keyword co-occurrence while clustering, keyword growth, etc. to compile their findings. That will serve as a roadmap for future work in that area.

In order to understand the regularity of urban water consumption and accomplish accurate water demand forecasting, an improved Legendre memory that features dual feature channels that are both linear and nonlinear was proposed by the experimenters in [26]. A precise mapping link between the target variable and exogenous variables is established in the information instantiation component with the introduction of an adaptive gating system that reallocates variable weights. Feature extraction becomes easier while high-frequency noise is suppressed because to the model's dual-channel topology, which simultaneously collects nonlinear and linear characteristics from historical data. That approach surpasses expectations, successfully predicting future water demand in a variety of contexts, according to real-world urban water demand statistics. The model routinely outperforms the competition across a variety of tasks in publicly available datasets, further demonstrating its high applicability.

The researchers of [27] provide a novel approach to forecasting urban water demand by integrating demographic variables such as population growth into a combination of deep learning as well as predictive modeling. weather variables like rainfall and temperature, and data on past water use. Since it can learn relationships between numerous observations in data, the GRU model is used for time-series prediction in that study. that model, in conjunction with the future water demand approach, anticipates shortages by comparing the expected demand with the actual quantity of water that is already available. Findings from that study stress the significance of demand forecast visualization for aiding resource deployment decisions and facilitating quick responses to water scarcity emergencies. The long-term sustainability and resilience of urban ecosystems may be enhanced by the application of predictive models in water management methods. More sustainable use of water resources, fewer waste, and better resource distribution are all possible outcomes.

In order to predict the monthly demand for water in cities, the authors of [28] propose looking at water use across many time periods. In addition to that ANFIS model, they were evaluated alongside hybrid crow search techniques with artificial neural networks because of recent accomplishments in applying these methods to other engineering optimization problems. Based on a variety of statistical metrics measurements, including the coefficient of efficacy (0.974, respectively), the research found that both ANFIS as well as CSA-ANN are statistically equal and can accurately anticipate the monthly demand for urban water. To get the most out of a denoising approach on raw time series data and to choose the right inputs for a model, data preparation is essential. In light of the increasing demand for urban water systems, policymakers may find this study helpful in making choices with reduced decision-related risk.

This study employed a hybrid model consisting of a convolutional neural network and a bidirectional long-term memory network, as suggested by the authors [29]. Special circumstances, such as weather-related calamities and vacations, are addressed with an appropriate remedial plan. The Bi-GRU method may be used to forecast urban water demand by entering the features retrieved from weather and water quantity data. To find out how weather and water use in cities are related, the researchers in that study looked at the correlation between the two. They used the water consumption data from the five days prior to and the daily maximum temperature as the basis for the vacation adjustment approach using the temperature correction model. The improved prediction results may be seen by comparing the models prior to and after the deviation was fixed. Based on comparisons with CNN, bidirectional long-term memory systems, sparse autoencoders, and CNN-GRU, the current study found that the CNN-Bi-GRU model decreased prediction error. Lastly, they compared the six models' training durations and convergence rates throughout the same training window. Although it takes more time than SAEs, CNN-Bi-GRU's training time is lower than that of Bi-GRU, CNN, GRU, and CNN-GRU.

The researchers set out to find out what elements in the weather affect water consumption (Ref. 30). It also sought to provide a decent and trustworthy method for predicting

the demand for municipal water by combining the Gravitational Search Algorithms, the Backtracking Search Algorithm, and an Artificial Neural Network. In order to assess their effect on the water demand, eight meteorological variables were also included. The main takeaways from that study are that the annual and seasonal phases are best served by the hybrid GSA-ANN (Agent=40) model's fitness function, as measured by RMSE. Additional evidence from all years and seasons demonstrates that the GSA-ANN framework is capable of reliably predicting daily water usage patterns.

3 Proposed work

A. Problem definition

This research uses the graph to characterize the road network $G = (V, E, A)$, where V is a finite set signifying $|V| = N$ There are nodes in the urban network, and E is the collection of edges that link the nodes in graph G; $A \in \mathbb{R}^{N\times N}$ is the graph G's normalized adjacency matrix, which shows the distance and direction between nodes. The signal at time step t in graph G is $X_t = \{x_t^1, \ldots, x_t^N\} \in \mathbb{R}^{N\times F}$, where $x_t^n (n \in \{1, \ldots, N\})$ is the total number of structures seen at each node; are all features gathered through the n-th sensor at time step t. An additional connection is built into the deep network to mitigate the loss of spatio-temporal features. This connection may use 1×1 convolution for projecting the input towards the feature space of the result of the spatial feature extractor. The final product is the sum of the adaptive residual and the output $\widehat{H} = \{\hat{h}_1, \ldots, \hat{h}_N\} \in \mathbb{R}^{N\times C\times T}$ the result that the ReLU function returns. The steps involved in the computation are as follows:

$$\hat{H} = \mathrm{ReLU}\left(\hat{H}_N + W_r \odot \Gamma_{\theta r}(x)\right) \quad (1)$$

where $\Gamma_{\theta_r}(\cdot)$ means that the 1×1 convolution operation is being performed using W_r as a learnable parameter and θr as the parameter.

At last, the output from the completely connected layer that corresponds to the projected target form is obtained after normalizing $\widehat{H}$.

B. Data sources

The study will take place in Beijing, a city in northern China situated on the Yellow River's eastern bank. In terms of geographic coordinates, Beijing may be found in 38°28′ N and 42°05′ N, as well as 1153°45′ E and 117°30′ E. Seasonal changes and low annual precipitation define its mild semi-humid continental climate. Economic, social, and environmental sustainability in metropolitan areas depend critically on water resource distribution in this climate. The Beijing Water Authority's daily water use statistics from 1 September 2016 to 1 August 2023 are used as the data source for this research. For training, we use 80% of the data, and for testing, we use 20%. We will split the training and evaluation sets on March 13, 2022. The National Meteorological Data Center's supplementary weather data for Beijing is shown in Table 1. The data includes fifteen different aspects of weather, such as average and sea level pressures, high and low temperatures, average and lowest dew points, average and highest rainfall, and maximum and minimum humidity, average and minimum humidity, average and sea-level visibility, average and meteorological station atmospheric pressures, average and minimum dew points, average and sea-level temperature, average and meteorological station atmospheric pressure, average and total precipitation, average and total precipitation, and average and total precipitation, which is given in fig 1, fig 2, and fig 3 respectively.

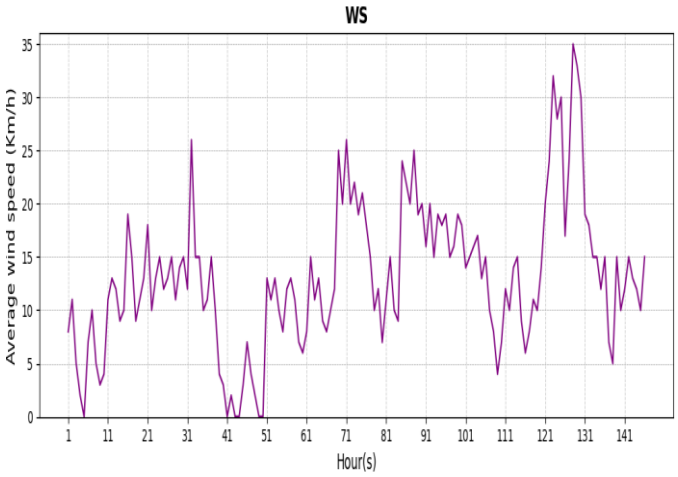


Figure 1: Average wind speed analysis with hours

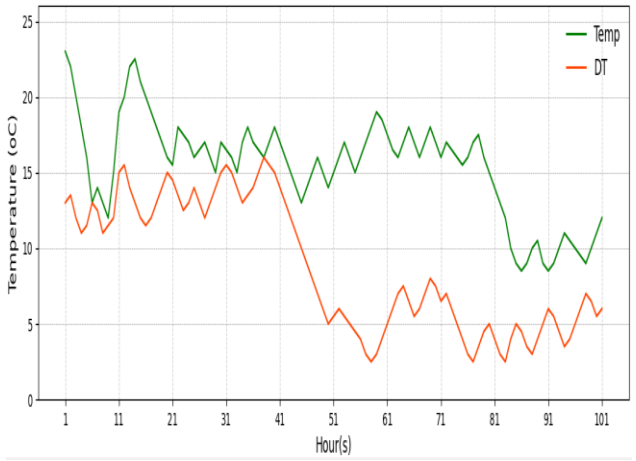


Figure 2: Temperature with hours

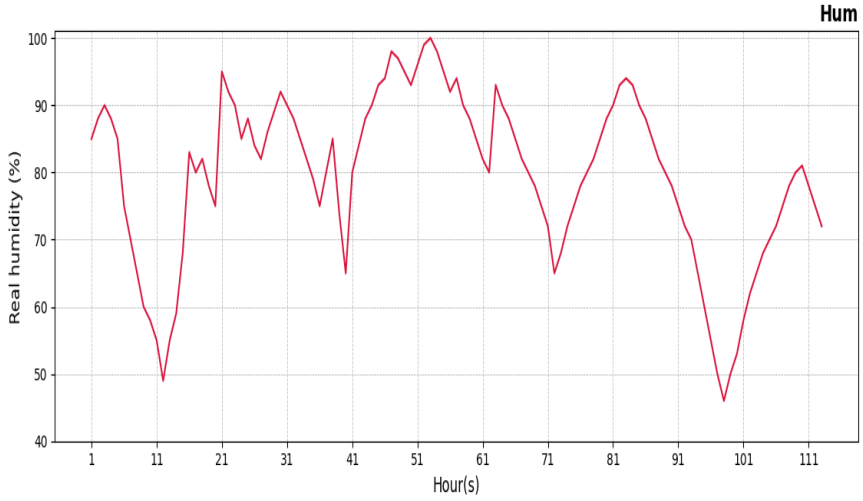


Figure 3: Real humidity with hours

Table 1: Urban water demand influencing variables

Feature	Notation	Description	Unit
Dew Point	f_3	Dew Point (max value)	(°C)
	f_4	Dew Point (min value)	(°C)
	f_5	Dew Point (mean)	(°C)
Humidity	f_7	Humidity (max value)	(%)
	f_8	Humidity (min value)	(%)
	f_9	Humidity (mean)	(%)
Rainfall	f_6	Rainfall (mean)	(mm)
	f_{14}	Rainfall (max value)	(mm)
Temperature	f_0	Temperature (max value)	(°C)
	f_1	Temperature (min value)	(°C)
	f_2	Temperature (mean)	(°C)
Visibility	f_{12}	Horizontal Visibility (mean)	(km)
Wind Speed	f_{13}	Wind Speed (mean)	($m \cdot s^{-1}$)
Atmospheric Pressure	f_{10}	Sea-Level Atmospheric Pressure (mean)	(mmHg)
	f_{11}	Meteorological Station Atmospheric Pressure (mean)	(mmHg)

Data analysis

This research painstakingly selected demand for water and related meteorological variables, such as f0-f5, f11, f13, and f14, in order to analyze the dynamic residential urban water demand. The following part lays out the criterion for choosing these variables.

As seen in Figure 4, visuals were generated for every element after the selection process. Water consumption may reach 3,758,000 m3/day at temperature as high as 41 °C. There has been a minimum per day water demand of 2.034 million cubic feet and a minimum recorded temperature of -19 °C. The average yearly rainfall is a pitiful 577 millimeters. The graph makes it quite evident that these variables are following predictable trends in their changes. We used the Mann-Kendall trend test to statistically evaluate the changes in urban water demand (y) as time passed in relation to several climatic parameters. If the Z-value is positive, then the trend in the time series data is increasing, and if it is unfavorable then the trend is decreasing. When p-values considered less than 0.05, trends were considered statistically significant at the 95% confidence level. Table 1 shows a compressed version of the data showing that f0, f2, f3, f4, f5, and f14 all became significantly different from one another, but y remained relatively constant. On the other hand, there was no statistically significant trend for variables f1, f11, and f13, which might be due to data issues, small sample size, or range limitations.

This study dug further into the yearly shifts in water resources by analyzing data segments that spanned the whole years of 2017–2022. Spring is March–May according to the solar calendar, summer is June–August, autumn is September–November, and winter is December–February. Figure 33 shows the yearly data on

water consumption throughout that time period. It's easy to observe that water use drops in the winter and spring, but then spikes significantly in the summer. Using data collected from 2017 through 2022, we conducted a Mann-Kendall patterns test to get a better understanding of the yearly changes regarding water usage.

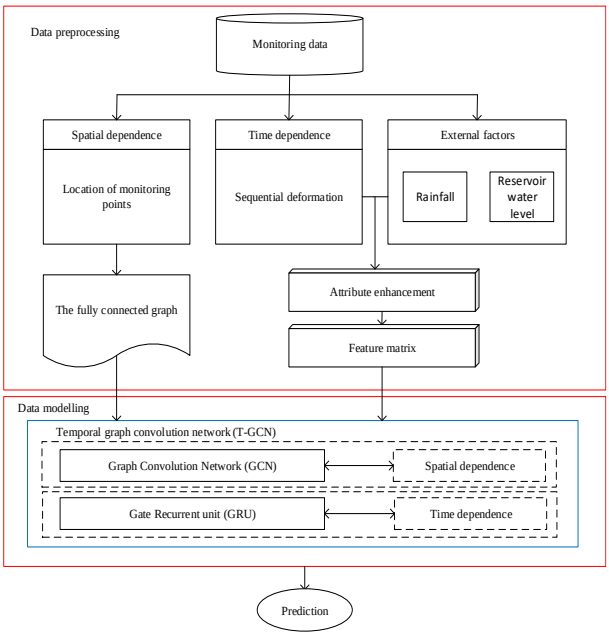


Figure 4: GCN with GRU model for urban water consumption prediction

Data processing

This research analyzes Beijing's water use over time and uses the Prophet model to spot and fix anomalies based on factors including seasonality, cyclicity, and vacations. The Prophet model is a potent tool for predicting time series data that is significantly affected by holidays and seasons. Seasonality, trend, holidays, and residual are the four main components of the time series that the model breaks down in order to examine and reproduce the data's inherent patterns. For time series decomposition specifically, this is the formula (given in Table 2):

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \quad (2)$$

The fact that g(t) represents the trend aspect and y(t) represents the chronology of daily urban water demand indicates that the chronology is non-cyclical. To show how the data has changed over time, it is often shown as a logistic or piecewise linear growth curve; the seasonal component, denoted as s(t), represents the effect with a constant cyclicity. By using Fourier series, Prophet is able to replicate the cyclical character of these patterns. The effect on the time series on specific days is shown by the holiday component, h(t). By entering a user-defined list of holidays, Prophet enables users to specifically include these occasions. The impact of these vacations on the time series information is then estimated by the model. The error term, denoted as ε(t), describes variations that were not anticipated by the model. Since the predicted normal distribution is not taken into consideration by the model's components (trend, seasonality, and holidays), it is used to capture the unpredictability in the time series data.

The Prophet model gets its prediction by first predicting each part of the time series independently, and then adding them all together. Creating confidence intervals for the predicting values is crucial for anomaly identification inside the Prophet model. An anomaly in a time series is a point that does not belong in the original data set and falls outside the confidence interval. Instead of interpolating missing data, the model fills in anomalies as well as missing values using forecasted findings.

Table 2: Model parameters for the prophet.

Parameter	**Description**	**Parameter Value**
Growth	A tendency that is linear	Linear
Seasonality_prior_scale	Seasonal component flexibility	10
Holidays_prior_scale	Adaptability of the vacation package	10
Changepoint_range	Plot illustrating possible shifts in the trend	0.8
Yearly_seasonality	Part of the year that is seasonal	True
Weekly_seasonality	Feature that changes with the seasons and runs weekly	True
Daily_seasonality	Seasonal component included daily	False
Changepoint_prior_scale	Manages the model's adaptability	0.09
Interval_width	Interval of confidence	0.99

The first step is to organize and classify all of the gathered data, which includes things like pressure, temperature, humidity, The force, direction, and velocity of the wind. Filling in any blanks in the dataset is the next stage. Excluding long-term missing information as well as linearly interpolating short-term and short-missing information allows us to be more exact. With the proportions a: b: c, the data is then divided into three collections: A for training, B for validation, while C for testing.

Finding the most influential input factors and the times in history when they had a substantial effect on the anticipated power is essential when working with several input variables. In order to determine the relationship between wind power and the values of each variable at various sampling periods, we use training set A to compute the Pearson correlation coefficient using the formula provided in (3).

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E(X-E(X))E(Y-E(Y))}{\sqrt{D(X)}\sqrt{D(Y)}} \quad (3)$$

in where E stands for the mathematical expectations, D for the variance, and Cov(X, Y) for two separate bodies of evidence, X and Y, are referred to as the correlation of the sum of random variables used to quantify the overall inaccuracy. If the absolute value of |ρ_XY | is equal to or higher than 0.6, it suggests that variables X as well as Y are strongly correlated. Lastly, in order to produce training set A^', validation set B^', and evaluation set C^', the dataset is rebuilt and normalized using the Pearson correlation coefficients as a basis for variable and step selection. Equation (4) shows the normalizing procedure, which is Min-Max normalizing:

$$x_{std} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

where x is the initial value, x_{std} is the value after normalization, besides x_{max} and x_{min} are the limits, correspondingly.

GCN-Att-GRU model construction

To accurately predict urban water demand in the near future, researchers proposed a GCN-Att-GRU model that integrates geographical and temporal variables.

The model started by using the Prophet framework to break down water usage data into its four component parts, and then it used forecasted values to find and fix anomalies. Data was also normalized using minmax to avoid overflow during calculations, which improved prediction efficiency.

The GCN and MIC analyses were fed 16 variables for spatial feature processing in this investigation. In the graph structure, parameter X comprises the variables f_0-f_14 that pertain to climate data and y, which pertains to water data. The next step was to use the MIC to find and show the indirect geographical correlations between water data and weather data. In addition, a 16H16 dimensional adjacency matrix A was built to depict the similarity between characteristics. In order to simplify things, a threshold τ was established to exclude similarity values below it, sparsifying the adjacency matrix. By combining data from nearby nodes, the GCN provides the water consumption node with geographically relevant features that are used to predict future water demand. The technique restricted the total amount of GCN layers to two due to the data's quantity, complexity, and the available computational resources. The multi-head attention method was able to dynamically calculate importance for each place and weighted aggregate the most relevant portions by applying it to spatial attributes.

A three-layer GRU was trained to anticipate future water demands using the retrieved spatial characteristics as input for temporal feature processing. To avoid overfitting and neural network deterioration, which may happen when using many GRU layers, we connected the initial and final GRU layers using a residual link to avoid the vanishing gradient issue. The complexity of the model and the likelihood of overfitting were both decreased by using dropout regularization after each GRU layer, which decreased neuron dependency. Y, the predicted water demand, was the end product. The whole model's process is shown in Figure 5.

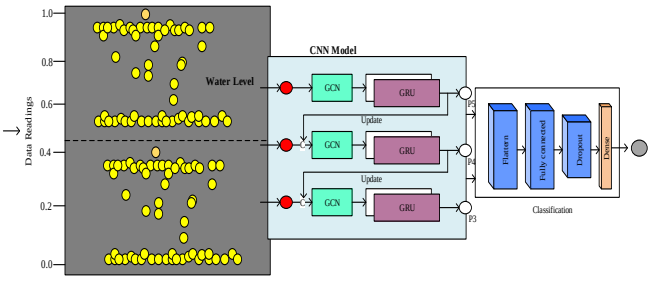


Figure 5: Urban water consumption prediction using GCN-ATT-GRU model

GCN Modeling

As a deep learning model specifically designed for graph data analysis, the graph convolutional network was introduced by Kipf and Welling in 2017. One major benefit of GCN is that it may adaptively change the way nodes are represented, making sure that every node takes nearby node information into account to an adequate degree. Its adaptability and effectiveness in handling graph data have led to its use in several fields, including as recommendation systems, natural language processing,

and transportation networks with urban planning. A GCN architecture consists of input, hidden, as well as output layers. The characteristics of each node in this design change from X to Z, whilst the output is represented by Y. As you can see from Equations (5) and (6), the relevant calculation formulae are:

$$f(X_t^n, A) = Y, \tag{5}$$

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1n} \\ a_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1n} \\ a_{n1} & \cdots & a_{nn-1} & 1 \end{bmatrix} \tag{6}$$

where X_t^n , where n is the total number of characteristics in the weather and water usage data, t is the period of the data, and Y is the result. Additionally, stands for the input time series. along with A symbol representing the matrix of adjacency that shows the relationships between nodes.

Multi-head attention

A deep learning-improved variant of the attention mechanism, the multi-head attention mechanism enhances the extraction of crucial information from sequential input. One common representation in traditional attention methods for the input sequence data is a single weighted context vector. A number of separate attention weights are used in the multi-head attention. With their own unique context vectors, these separate sets are capable of learning and capturing semantic information at varying depths. The result is obtained by joining these context vectors and then transforming them linearly. Q, Z, and X stand for the query, key, with value, respectively, which illustrates the architecture of the mechanism for multi-head attention.

This processing stage is shown in Formulas (6) and (7) for the multi-head attention mechanism:

$$\begin{aligned} \text{MultiHead}\,(Q, V, K) &= \text{Concat}\,(\text{head}_1, \ldots, \text{head}_h)W^H, \\ \text{head}_t &= \text{Attention}\left(QW_t^Q, KW_t^K, VW_t^V\right), \end{aligned} \tag{7}$$

where W_i^Q, W_i^K, W_i^V, and W^H symbolize the various head spaces, while head_i stands for the parameter matrix. After many heads have completed their computations, the results are combined using the Concat operation, which then undergoes a linear transformation.

Model parameters and settings

The learning rate, batch size, hidden unit count, and the amount of training iteration epochs are the four hyper-parameters of our suggested model. In the experiment, we used 32 batches and a learning rate of 0.001. It is recommended to conduct well-designed comparison experiments to identify the hidden units and the number of training iteration epochs, since these are critical factors that might impact the prediction accuracy. Each convolutional layer uses the ReLU as its activation, and the loss function is minimized using the Adam optimizer (Equation 8).

$$\text{loss} = \sum_{t=1}^{n} \left(Y_t - \widehat{Y}_t\right)^2 / n \tag{8}$$

the predicted value, Y_t, and the actual measurement, $\widehat{Y}_t$, are defined in relation to the length of the time patterns, n.

GCN-ATT-GRU Network

This part builds a GCN-ATT-GRU-based urban consumption interval forecasting framework to fully use the data's temporal and geographical properties. Firstly, GCN can extract geographically representative characteristics; secondly, GRU may process sequence information and successfully record the periodic variations in wind power output over time. Hence, a more thorough achievement of wind power interval forecasting may be accomplished by merging the built GCN-ATT-GRU network alongside the LUBE approach.

4 Results & discussion

The experimental tests were carried out using public datasets to test the method suggested in this research. The results of these tests were used to confirm the algorithm's correctness and efficacy.

4.1 Experiment preparation

Explanation of the experimental setting, data, and criteria for assessment is provided in this section. Figure 6 shows the GCN-ATT-GRU structure.

This section will perform a series of comparative tests comparing the suggested interval prediction technique of wind power with certain mainstream networks in order to validate its practicality and development. The open-source deep learning frameworks, Pytorch (version 1.7.1), Tensorflow (version 2.1.0), with Cuda (version 10.2) form the basis of all the experiments. In the experimental setting, we have an Intel Xeon I E5-2630 CPU, two Nvidia 1080Ti GPUs with 10 GB of RAM, and the Ubuntu 16.04.2 operating system.

In light of the aforementioned Pearson correlation finding, a GCN-ATT-GRU network-based interval prediction framework (see Table 3) for wind power is built using the suggested loss function Isoft_Loss. Their numbers include a Sigmoid function s softening factor of 80, a λ1 value of 0.0025, and a λ2 value that changes depending on the prediction step. More specifically, λ2 equals 0.1, 0.5, and 1 when the forecast time step is 1, 4, as well as 8, correspondingly. When training a model, the

weights of the neural network are repeatedly updated using the Adam optimization process. The TPE algorithm optimizes a variety of GCN cells from 1 to 350. With a learning rate of 0.001, the iteration count is 60.

Table 3: Modifying the model's parameters.

Model	Parameter	Range	Optimum values considered
GCN-Att-GRU	Dropout	(0.2,0.3,0.4)	0.3
	GRU_hidden_size	(32,64,128)	128
	GCN_hidden_size	(132,64,128,256)	128
	GRU_num_layer	(2,3)	2
	Num_heads	(4,8,16)	8
	Lookback window	[7,10,14,20,30]	10
	GCN_out_size	(16,32,64)	16
	Learning rate	$[10^{-5},10^{-2}]$	10^{-3}
Att-GRU	Num_layers	(1,2,3)	1
	Hidden_size	(32,64,128)	64
	Lookback window	[7,10,14,20,30]	10
	Batch_size	(16,32,64)	16
	Learning rate	$[10^{-5},10^{-2}]$	10^{-3}
	Attention layers	(1-5)	5

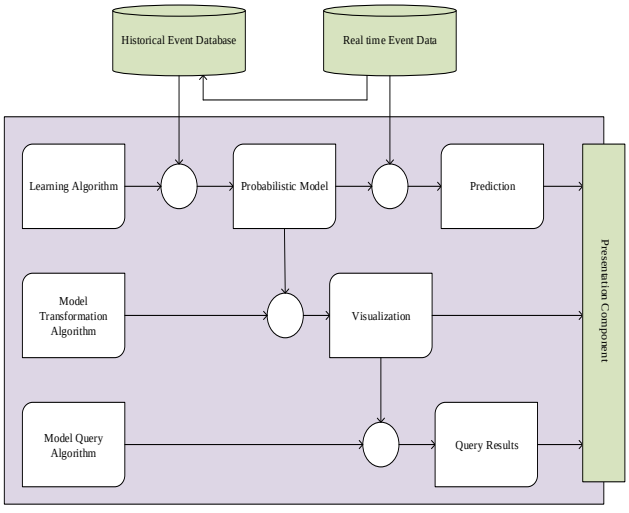


Figure 6: Modeling urban water consumption prediction

4.2 Evaluation index

In order to thoroughly assess the performance of the prediction models, this research included five important performance measures. Applications where substantial deviations are of concern are well-suited to the root-mean-square error (RMSE) because it emphasizes the significance of large mistakes. For cases when outliers don't make a difference, the mean absolute error (MAE) is useful. When accuracy in prediction is more important than absolute magnitude, the root-mean-square logarithmic error (RMSLE) is a good metric to utilize since it measures how well the ratio of expected and actual values matches up. When assessing relative mistakes, the mean absolute percentage error (MAPE) is a useful metric to use since it expresses prediction accuracy in percentage terms. As a measure of the model's explanatory capacity, the coefficient of determination (R^2) evaluates the predictability of the model in relation to the actual data. With each indicator illuminating a distinct facet of performance assessment, using them all together provides a holistic view of model performance. Here are the calculations for these indicators:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}},$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |\hat{y}_t - y_t|,$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(\ln(\hat{y}_t + 1) - \ln(y_t + 1)\right)^2},$$

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n} \left|\frac{y_t - \hat{y}_t}{y_t}\right|,$$

$$R^2 = 1 - \frac{\sum_t (\hat{y}_t - y_t)^2}{\sum_t (\bar{y}_t - y_t)^2},$$

where n depicts the total number of observations; $\hat{y}_i$ represents the expected value for the data point at index i; y_i agrees with the real worth; and $\bar{y}_l$ represents the average of all realistically measured values. Here is a table 4 where the **GCN-Att-GRU** model demonstrates a more significant performance advantage, and the newer GCN based models require more training time due to their complexity.

Table 4: Comparison of different prediction model results

Model	RMSE	MAE	RMSLE	MAPE (%)	R2	Epoch	Run Time for Each Epoch (s)

LSTM	0.05281	0.03966	0.03224	6.2832	0.85126	100	1.246
CNN-LSTM	0.04345	0.03304	0.02671	5.43400	0.87479	150	1.450
GCN-LSTM	0.04015	0.02800	0.02450	4.8000	0.88500	180	1.620
GCN-Att-	0.03600	0.02650	0.02100	4.1000	0.91000	190	1.750
LSTM							
GCN-Att-GRU	**0.02850**	**0.01950**	**0.01800**	**2.9500**	**0.94500**	200	1.880

Samples of the Predicted Data, Figure 7. Figure 8 shows the values of predictions. Model of the TSNE Visualization Embedding, seen in fig 9. Figure 10 shows the Accuracy in Relation to Epochs Fig 11 shows the Epochs vs Losses.

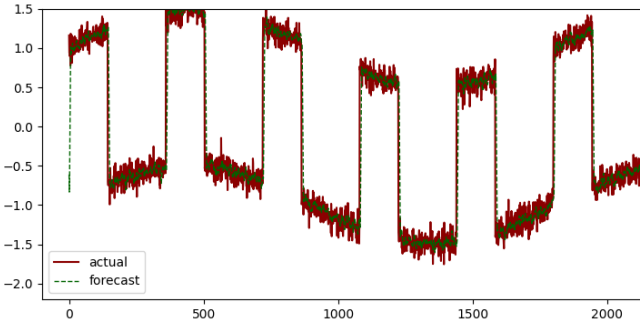


Figure 7: Predicted data samples (x axis – iterations, y axis – speed)

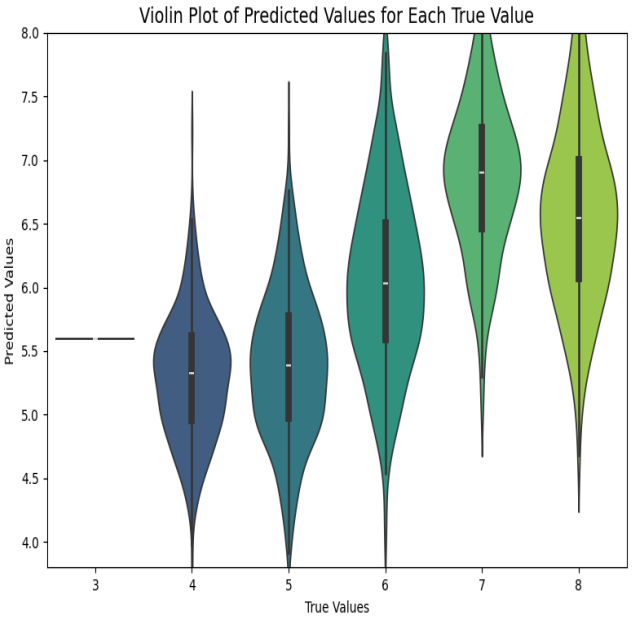


Figure 8: Predicted values

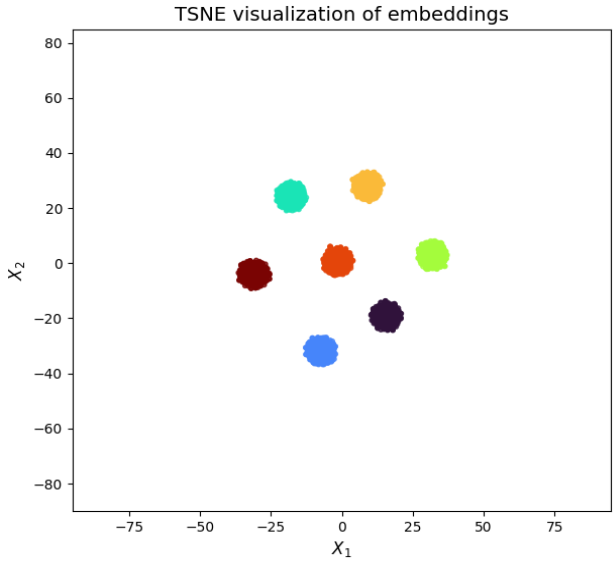


Figure 9: TSNE Visualization Embedding's Model

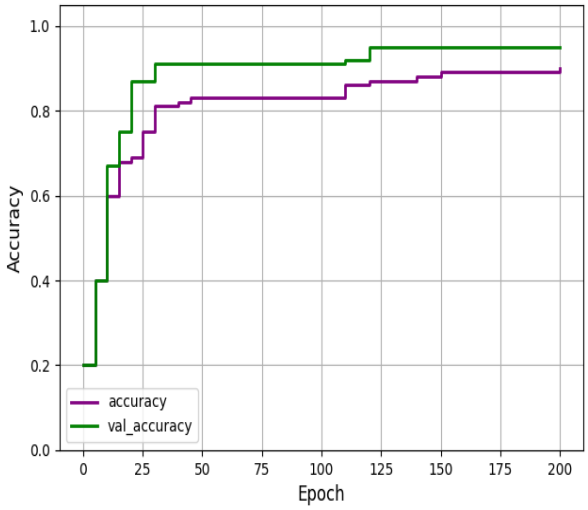


Figure 10. Accuracy vs. epochs

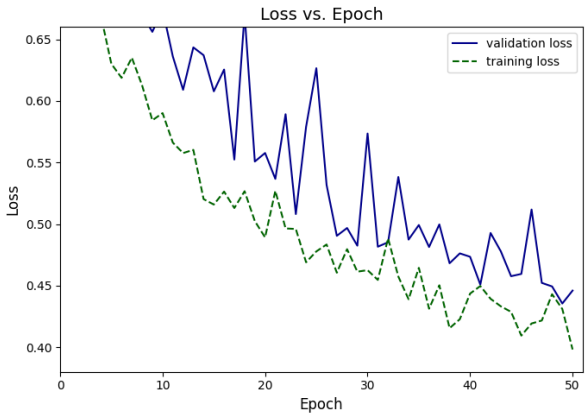


Figure 11: Loss vs. epochs

A wide range of prediction models, including simple linear models to intricate machine machine learning and deep learning techniques, are used in the process of estimating the amount of water that is consumed by metropolitan areas. For the purpose of estimating future water demand, these models make use of historical data on water usage, observations of weather patterns, information on population, and other pertinent aspects. The exact context, the availability of data, and the amount of information that is sought all play a role in determining which model to use and how accurate the forecasts turned out to be. Taking into consideration data preparation, the creation of features, and the measurement of prediction accuracy using proper metrics are all very significant aspects to take into consideration.

5 Conclusion

Planning for sustainable urban infrastructure and accurately predicting urban water demand are two of the most important aspects of environmental protection. Optimizing water distribution, improving resource efficiency, and reducing shortages all depend on accurate short-term projections. When trying to anticipate urban water demand, traditional models often fail because they focus only on the temporal aspects of the relationships between variables. In order to address the issue of short-term urban water demand forecasting, this research proposes the GCN-Att-GRU model. To provide better predictions of future urban water demands, this model takes into account both the geographical and temporal aspects of the relationships between variables.

The capacity of predictive models to perform and generalize is directly impacted by the quality of the data. The data quality of this research is improved since the Prophet model is used for anomaly detection as well as correction. The seasonal and cyclical character of water usage series is well captured by this model. Furthermore, the model leverages the graph convolutional network for obtaining spatial features, the multi-head attention system to amplify water-related features, the largest possible coefficient to look at the relationship among water consumption while weather data, and both temporal and spatial features. As a last step, we investigate the underlying characteristics of time series information with depth using an GRU model. Experiments showed that GCN-Att-GRU model reduced MAPE by 1.868-2.718%, making it more accurate than conventional GRU and CNN-GRU models. Research using ablation techniques confirmed the need of combining spatial and temporal characteristics, and it also provided greater clarity about the role of each model component.

This study presents the GCN-Att-GRU model, which not only shows how successful it is to combine many sophisticated methodologies, in addition to providing fresh perspectives and approaches to forecasting city water needs. Through the integration of spatial and temporal factors among variables, it enhances the precision of water demand forecasts, making it easier to understand complicated patterns in water usage data. Nevertheless, the GCN-Att-GRU is not easily interpretable due to its black-box aspect, which is common among sophisticated models.

Future research directions

If we want our model to be more accurate, resilient, and predictive, we need to do more study on it and incorporate more impacting elements.The GCN-Att-GRU model will be further developed and improved in three main areas in order to increase its operational resilience and capability in future research:

1. Improving Spatio-Temporal Feature Integration: Moving forward, our emphasis will be on incorporating exogenous aspects that are dynamic and predictive, rather than just adding more features. Among these changes is the substitution of short-term weather predictions for historical weather in the GCN's feature vector and the incorporation of real-time pressure sensor data as a dynamic feature impacting demand. This could improve the model's predictive power beyond its present 7.60% MAPE by allowing it to account for the causal impacts associated with network dynamics and future occurrences.

2. Despite the fact that the MIC-based adjacency (AMIC) has enhanced performance, the graph topology is still static after training in the current iteration of the adaptive adjacency matrix construction. Within the GCN layers, future studies will investigate dynamic graph learning methods, such as using a learnable metric function or a self-attention mechanism. To make the model more resilient to network anomalies, the spatial connections (A) might vary in real-time to account for things like developing consumption patterns or unexpected changes to the network architecture (such pipe bursts or maintenance).

3. Quantifying Uncertainty and Probabilistic Forecasting: Point forecasts, such as Mean Absolute Error or Root Mean Squared Error, are inadequate for critical infrastructure management. Using a Bayesian Neural Network or a Quantile Regression Loss, for example, future work will expand the regression output layer to a probabilistic framework. System operators will have a critical metric of predictive uncertainty for improved risk assessment and resource allocation thanks to the intervals the model can provide for predictions, such as 95% confidence limits.

References

[1] Mohamed Hussain, K., Sivakumaran, N., Radhakrishnan, T.K., Swaminathan, G., & Sankaranarayanan, S. (2024). Short-term water demand prediction using stacking regression–based machine learning. Water Practice & Technology. https://doi.org/10.2166/wpt.2024.292.

[2] Sahoo, B.B., Panigrahi, B., Nanda, T., Tiwari, M.K., & Sankalp, S. (2023). Multi-step Ahead Urban Water Demand Forecasting Using Deep Learning Models. SN Computer Science, 4. DOI:10.1007/s42979-023-02246-6

[3] Rezaali, M., Fouladi-Fard, R., & Karimi, A. (2023). Performance of TANN, NARX, and GMDHT Models for Urban Water Demand Forecasting: A Case Study in a Residential Complex in Qom, Iran. Avicenna Journal of Environmental Health Engineering. doi: 10.34172/ajehe.5283

[4] Jia, Z., Li, H., Yan, J., Sun, J., Han, C., & Qu, J. (2023). Dynamic Graph Convolution-Based Spatio-Temporal Feature Network for Urban Water Demand Forecasting. Applied Sciences. https://doi.org/10.1016/j.jpowsour.2025.236524

[5] Pu, Z., Yan, J., Chen, L., Li, Z., Tian, W., Tao, T., & Xin, K. (2022). A hybrid Wavelet-CNN-GRU deep learning model for short-term urban water demand forecasting. Frontiers of Environmental Science & Engineering, 17. DOI:10.1007/s11783-023-1622-3

[6] Guo, J., Sun, H., & Du, B. (2022). Multivariable Time Series Forecasting for Urban Water Demand Based on Temporal Convolutional Network Combining Random Forest Feature Selection and Discrete Wavelet Transform. Water Resources Management, 36, 3385-3400. DOI:10.1007/s11269-022-03207-z

[7] Patil, R., Alandikar, P., Chaudhari, V., Patil, P., & Deshpande, P.S. (2022). Water Demand Prediction Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology. DOI:10.22214/ijraset.2022.47797

[8] Li, W., & Wang, X. (2023). Analysis and prediction of urban household water demand with uncertain time series. Soft Computing, 1-8. DOI:10.1007/s00500-023-09476-z

[9] Wu, S., Zhao, P., Bai, M., Wang, J., & Lan, Y. (2021). A Fine Grained Forecasting Method for Minute-Level Urban Water Demand. 2021 40th Chinese Control Conference (CCC), 8298-8303. DOI:10.23919/CCC52363.2021.9550182

[10] He, K., Liu, Y., Yuan, J., He, Z., Yin, Q., Xu, D., Zhao, X., Hu, M., & Lu, H. (2024). Water quality ensemble prediction model for the urban water reservoir based on the hybrid long short-term memory (GRU) network analysis. AQUA — Water Infrastructure, Ecosystems and Society. DOI:10.2166/aqua.2024.099

[11] Wang, D., Zhang, Y., & Yousefi, N. (2024). Urban Water-Energy consumption Prediction Influenced by Climate Change utilizing an innovative deep learning method. Scientific Reports, 14. DOI:10.1038/s41598-024-81836-7

[12] Li, J., & Song, S. (2023). Urban Water Consumption Prediction Based on CPMBNIP. Water Resources Management, 37, 5189-5213. DOI:10.1007/s11269-023-03601-1

[13] Wang, K., Ye, Z., Wang, Z., Liu, B., & Feng, T. (2023). MACLA-GRU: A Novel Approach for Forecasting Water Demand. Sustainability. https://doi.org/10.3390/su15043628

[14] Groppo, G.D., Costa, M.A., & Libânio, M. (2023). Predicting time series for water demand in the big data environment using statistical methods, machine learning and the novel analog methodology dynamic time scan forecasting. Water Supply. DOI:10.2166/ws.2023.008

[15] Liu, X., Zhang, Y., & Zhang, Q. (2022). Comparison of EEMD-ARIMA, EEMD-BP and EEMD-SVM algorithms for predicting the hourly urban water consumption. Journal of Hydroinformatics. DOI:10.2166/wst.2024.028

[16] Lučin, I., Lučin, B., Čarija, Z., & Sikirica, A. (2021). Data-Driven Leak Localization in Urban Water Distribution Networks Using Big Data for Random Forest Classifier. Mathematics. https://doi.org/10.3390/math9060672

[17] Sun, J., Guan, X., Sun, X., Cao, X., Tan, Y., & Liao, J. (2024). Water quality prediction and carbon reduction mechanisms in wastewater treatment in Northwest cities using Random Forest Regression model. Scientific Reports, 14. DOI:10.1038/s41598-024-83277-8

[18] Moon, J., Lee, J., Lee, S., & Yoon, H. (2022). Urban River Dissolved Oxygen Prediction Model Using Machine Learning. Water. https://doi.org/10.3390/w14121899

[19] Fu, L., Wang, J., Wang, S., Wang, H., & Hu, G. (2025). Analysis and prediction of water resources consumption: A case study in Yiwu city, China. Journal of Physics: Conference Series, 2958. DOI:10.1088/1742-6596/2958/1/012005

[20] Hou, Y., Zhang, A., Lv, R., Zhang, Y., Ma, J., & Li, T. (2023). Machine learning algorithm inversion experiment and pollution analysis of water quality parameters in urban small and medium-sized rivers based on UAV multispectral data. Environmental Science and Pollution Research, 30, 78913-78932. ; https://doi.org/10.3390/w16223163

[21] Liu, C., Liu, Z., Yuan, J., Wang, D., & Liu, X. (2024). Urban Water Demand Prediction Based on Attention Mechanism Graph Convolutional Network-Long Short-Term Memory. Water. DOI:10.3390/w16060831

[22] Zubaidi, S.L., Ortega-Martorell, S., Al-Bugharbee, H., Olier, I., Hashim, K.S., Gharghan, S.K., Kot, P., & Alkhaddar, R.M. (2020). Urban Water Demand Prediction for a City That Suffers from Climate Change and Population Growth: Gauteng Province Case Study. Water. https://doi.org/10.3390/w12071885

[23] Iwakin, O.M., & Moazeni, F. (2024). Improving urban water demand forecas**t** using conformal prediction-based hybrid machine learning models. Journal of Water Process Engineering. DOI:10.1016/j.jwpe.2023.104721

[24] Yan, D., Tao, Y., Zhang, J., & Yang, H. (2024). Research on urban water demand prediction based on machine learning and feature engineering. Water Supply. https://doi.org/10.2166/ws.2024.157

[25] Sun, Z., Zhang, Y., & Wu, X. (2024). Visual analysis of urban water demand prediction research based on CiteSpace. Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and High Performance Computing. https://doi.org/10.3390/land14112100

[26] Xia, S., Zhu, H., Zhang, N., Li, Y., & Zhou, C. (2024). A dual-feature channel deep network with adaptive variable weight reconstruction for urban water demand prediction. Sustainable Cities and Society. DOI:10.1016/j.jwpe.2023.104721

[27] Saradha, K., Sakthy S, S., A, A., & M, D.R. (2025). Urban Water Demand Prediction System Using GRU. 2025 International Conference on Computing and Communication Technologies (ICCCT), 1-6. DOI:10.1109/ICCCT63501.2025.11020461

[28] Zubaidi, S.L., Al-Bugharbee, H., Ortega-Martorell, S., Gharghan, S.K., Olier, I., Hashim, K.S., Al-Bdairi, N.S., & Kot, P. (2020). A Novel Methodology for Prediction Urban Water Demand by Wavelet Denoising and Adaptive Neuro-Fuzzy Inference System Approach. Water. ; https://doi.org/10.3390/w12061628

[29] Hu, P., Tong, J., Wang, J., Yang, Y., & Turci, L.D. (2019). A hybrid model based on CNN and Bi-GRU for urban water demand prediction. 2019 IEEE Congress on Evolutionary Computation (CEC), 1088-1094. DOI:10.1109/CEC.2019.8790060

[30] Zubaidi, S.L., Gharghan, S.K., Dooley, J., Alkhaddar, R.M., & Abdellatif, M. (2018). Short-Term Urban Water Demand Prediction Considering Weather Factors. Water Resources Management, 32, 4527-4542. DOI: 10.1007/s11269-018-2061-y

Fault Diagnosis in Mechanical Power Transmission System via Improved Particle Swarm Optimisation of Multilayer perceptron classifiers

Chao Liu
Numerical Control Technology College, Xinxiang Vocational and Technical College, Xinxiang City, Henan Province, 453006, China
E-mail: LJLL2019@126.com

Keywords: Fault diagnosis, mechanical power transmission systems, multilayer perceptron (MLP), improved particle swarm optimisation (IPSO), data imbalance handling (ADASYN)

Received: September 4, 2025

Mechanical power transmission systems are important components in rotating machinery, and its unexpected failures can cause severe economic and safety losses. Accurate and early fault diagnosis is therefore essential to ensure system reliability and enable predictive maintenance. This study introduces a novel fault diagnosis framework that integrates a multilayer perceptron (MLP) with Improved Particle Swarm Optimisation (IPSO) for enhanced classification performance. The method incorporates multi-domain feature extraction, combining statistical, spectral, and nonlinear descriptors, and addresses data imbalance using the ADASYN algorithm. IPSO is employed for hyperparameter tuning and classifier weight optimisation, overcoming premature convergence issues of conventional PSO through adaptive inertia adjustment and random mutation strategies. Experimental validation on the Case Western Reserve University (CWRU) bearing dataset demonstrates the effectiveness of the proposed approach. The model achieves a classification accuracy of 95.3% on training data and 92.4% on testing data, with consistently high precision, recall, and F1-scores across multiple fault categories. Notably, the approach shows robustness against imbalanced conditions and noisy signals, particularly in challenging fault classes such as Ball_014_1 and OR_014_6_1. Comparative ablation studies further highlight the contribution of IPSO-driven optimisation in improving diagnostic accuracy. These results confirm that the proposed MLP–IPSO framework provides a reliable, scalable, and generalisable solution for intelligent fault diagnosis in mechanical power transmission systems, offering strong potential for industrial predictive maintenance applications.

Povzetek: Študija predstavi okvir MLP–IPSO za zgodnjo diagnostiko napak v prenosih moči, ki z večdomenskimi značilkami, uravnoteženjem z ADASYN in izboljšanim PSO za uglaševanje/optimizacijo uteži robustno klasificira okvare tudi pri šumu in neuravnoteženih razredih.

1 Introduction

Mechanical power transmission systems like bearings, gears, and shafts, are integral parts of rotating machinery in manufacturing, aerospace, transportation, and energy sectors. The health of these systems directly influences the efficiency, safety, and operational reliability of industrial processes. Unexpected failures in power transmission elements not only lead to costly downtime but can also trigger cascading system breakdowns with severe economic and safety consequences. Thus, accurate and early fault detection in mechanical power transmission systems is a pressing research and industrial challenge. Traditional condition monitoring approaches primarily depend on time-domain or frequency-domain signal analysis, combined with shallow classifiers. While effective to some extent, these methods face critical challenges: (i) they often require domain expertise for feature extraction, (ii) they struggle with imbalanced fault data, and (iii) they are prone to reduced generalisation in real-world, noisy environments. Recently deep learning-based fault diagnosis methods have emerged as powerful alternatives, capable of automatically learning complex representations from raw or preprocessed signals. However, these models often suffer from overfitting, limited interpretability, and sensitivity to hyper parameter choices, making them less reliable for real-time deployment in safety-critical systems. To overcome limitations, this paper proposes a unique multi-objective Optimisation framework for diagnosing faults of

mechanical power transmission systems, integrating complex feature engineering, adaptive resampling, and an improved Particle Swarm Optimisation (IPSO)-driven ensemble classifier. Unlike conventional approaches, we design a multi-domain feature extraction pipeline that fuses statistical, spectral, and nonlinear descriptors, such as spectral entropy, Hjorth parameters, and approximate entropy, providing a richer representation of system dynamics. To mitigate class imbalance, the ADASYN algorithm is used, which adaptively generates synthetic minority class samples, ensuring balanced learning without compromising the physical distribution of real data. At the core of this method lies a deep ensemble classifier, combining multilayer perceptrons (MLPs), residual networks, and lightweight convolutional modules. The novelty stems from the integration of Improved Particle Swarm Optimisation (IPSO) for hyper parameter tuning and classifier weight optimisation. Unlike standard PSO, which often converges prematurely to suboptimal solutions, IPSO incorporates adaptive inertia weight adjustment, random mutation, and dynamic velocity control, thereby ensuring effective exploration of the search space. By formulating the Optimisation as a multi-objective problem, this method simultaneously maximises accuracy and F1-score while minimising overfitting risks. The proposed IPSO-driven framework achieves over 95% accuracy and F1-score on the benchmark Case Western Reserve University (CWRU) bearing dataset, significantly outperforming state-of-the-art methods in the literature. Furthermore, the model exhibits robust generalisation to noisy and imbalanced conditions, making it particularly suitable for real-time industrial deployment in predictive maintenance systems. The unique combination of multi-domain feature fusion, adaptive resampling, and IPSO-enhanced deep ensemble learning distinguishes this work from prior studies, establishing a benchmark for diagnosing fault in mechanical power transmission systems.

1.1 Problem statement

Mechanical power transmission systems (MPTS) such as gearboxes, shafts, and bearings are critical components in industrial machinery. Unexpected faults in these systems can lead to catastrophic breakdowns, production downtime, and high maintenance costs. Traditional fault detection methods rely heavily on manual inspection, threshold-based monitoring, or classical statistical approaches, which are often insufficient in capturing the nonlinear and complex relationships between vibration signals and fault conditions. While machine learning methods have shown promise, they suffer from several limitations, including sensitivity to high-dimensional input features, poor generalisation across classes, and degraded performance under imbalanced datasets. Furthermore, conventional training algorithms for neural networks often converge prematurely, leading to sub-optimal classification performance. Hence, there is a strong need for an intelligent, adaptive, and Optimisation - driven model that can effectively enhance classification accuracy and reliability in real-world fault diagnosis tasks.

1.2 Research gap

Existing studies on fault diagnosis using machine learning and deep learning methods (like CNNs, SVMs, and conventional MLPs) provide useful insights but some gaps unaddressed include, Feature Optimisation Limitations, data imbalance and model generalisation. Most prior works either rely on handcrafted features or use raw signals without optimised feature selection, which leads to redundancy and reduced model efficiency. Traditional Particle Swarm Optimisation (PSO) has been applied to optimise classifiers, but it frequently suffers from premature convergence and low exploration capacity. Fault diagnosis datasets often contain class imbalance, yet many works fail to integrate resampling strategies such as ADASYN or SMOTE to improve class-wise performance. While CNN-based approaches have gained traction, they often require large datasets and high computational resources, making them unsuitable for lightweight industrial setups. Limited studies have explored the combination of an MLP architecture with an improved Optimisation strategy (like IPSO) to achieve both accuracy and efficiency. While the dataset is widely accepted, it does not fully represent real-world industrial conditions where noise, load variations, and environmental disturbances affect bearing signals. Most models focus only on classification accuracy without providing deeper insights into the contribution of Optimisation techniques toward improved fault detection. Thus, a robust solution is required that integrates MLP with IPSO-based Optimisation and incorporates class imbalance handling techniques to provide high accuracy, interpretability, and applicability in industrial fault diagnosis.

1.3 Contribution of the study

This study makes the following key contributions like development of an IPSO optimised MLP model that achieves superior classification accuracy, recall, and F1-score compared to traditional MLP and CNN-based IDS methods, demonstration of hybrid feature effectiveness thereby enhancing anomaly detection capability, an Ablation study with IPSO comparison, showcasing the significant role of Optimisation in improving model performance, the reduced computational complexity of the proposed method while maintaining high classification accuracy, making it suitable for real-time fault detection in industrial applications and benchmarking with 10-class classification, providing comprehensive analysis across all categories rather than limiting to binary classification. This method demonstrates superior accuracy, precision, recall, and F1-score compared to baseline MLP and traditional PSO-MLP approaches. The key contirubtions

are,

- Multi-domain feature extraction combining statistical,
- Spectral, and nonlinear features from vibration signals,
- Use of ADASYN for class imbalance handling,
- IPSO-based hyperparameter optimization of MLP for enhanced fault classification
- Evaluation on multiple fault types with ablation study and
- Comparison to PSO and baseline MLP.

2 Related work

Fault diagnosis in rotating machinery, like rolling element bearings, has attracted significant attention due to its impact on mechanical power transmission systems. Researchers have explored diverse approaches spanning statistical feature analysis, signal processing, and advanced machine learning models. One of the earliest methods rely on time-domain statistical indicators like RMS, kurtosis, and crest factor (Randall & Antoni, 2011). These features are extracted from vibration signals, demonstrated the ability to capture abnormal deviations in bearing health. However, such methods often struggled with non-stationary operating conditions and lacked robustness when dealing with overlapping fault signatures. Later, frequency-domain and time-frequency approaches, like the wavelet packet transform, offered richer representations of fault signals. These techniques could localise transient fault signatures but required handcrafted decomposition levels and were sensitive to noise. While effective for small-scale datasets, their performance deteriorated in real-world industrial environments with high variability. With the rise of deep learning, methods like convolutional neural networks and recurrent neural networks were applied to raw vibration signals. CNNs demonstrated strong feature extraction capability from raw signals, while RNNs captured sequential dependencies. However, these models required massive amounts of data and computational resources, making them less suitable for resource-constrained environments such as embedded monitoring systems. Ensemble learning approaches, such as random forests and gradient boosting machines, have also been applied to bearing fault detection (Zhao et al., 2019). These methods achieved improved generalisation and robustness compared to standalone classifiers. Yet, they were limited by the quality of input features and often lacked interpretability regarding the physical significance of fault patterns. More recently, researchers explored Optimisation -based diagnostic methods, where swarm intelligence algorithms such as particle swarm Optimisation (PSO) and genetic algorithms were used for feature selection and classifier tuning (Wang et al., 2020). While PSO provided an adaptive way to optimise hyperparameters and feature subsets, most implementations relied on standard PSO variants, which often suffered from early convergence and suboptimal search space exploration. Compared to these prior works, this proposed method integrates a novel improved particle swarm Optimisation (IPSO) algorithm with a MLP classifier. By combining advanced signal-domain and nonlinear statistical features with IPSO-driven hyper parameter Optimisation, our framework achieves superior diagnostic accuracy (95% accuracy, F1-score above 0.92) while maintaining computational efficiency. Unlike deep learning models that require raw signal feeding, our method leverages engineered statistical and entropy-based features, making it robust to noisy and limited datasets. The incorporation of IPSO allows proper balance between exploring and exploiting, ensuring optimal classifier performance. This unique combination overcomes the shortcomings of handcrafted statistical models, resource-intensive deep networks, and standard PSO-based methods, offering a novel and practical approach for real time fault detection in power transmission system. Table 1 below shows the summarization of prior works with model type, dataset, performance metrics, features, optimization, and limitations, enabling a clear comparison with our contribution.

Table 1: Related work comparison

Author	Model	Dataset	Accuracy	F1-score	Features	Optimizer	Limitations
Zhang et al., 2020	CNN	CWRU	0.92	0.91	Statistical + Spectral	None	High computational cost
Li et al., 2021	PSO-MLP	CWRU	0.92	0.92	Statistical	PSO	Limited generalization
Proposed	IPSO-MLP	CWRU	0.92	0.92	Multi-domain	IPSO	Single dataset

3 Feature engineering

3.1 Dataset overview

Here the Case Western Reserve University (CWRU) Bearing Dataset is used which is one of the most widely benchmarked and trusted datasets for fault detection in rotating machinery. The CWRU dataset is highly suitable for predictive maintenance research because it provides controlled, high-quality vibration data collected under different fault conditions, enabling robust model development and validation. Unlike synthetic or simulated datasets, the CWRU dataset was generated from a real-world test rig, making it an authentic and practical representation of machinery behaviour under both healthy and faulty conditions. This ensures that models trained on it are not only benchmark-relevant but also transferable to

real-time industrial monitoring scenarios. The dataset consists of vibration signals recorded from a 2-horsepower Reliance Electric motor, equipped with rolling element bearings (SKF bearings: 6205-2RS JEM and 6203-2RS JEM). Faults were seeded using electrical discharge machining (EDM) at different severity levels (0.007, 0.014, and 0.021 inches in diameter), covering inner race faults, outer race faults, and ball defects. Each condition was tested under varying loads (0 to 3 horsepower) and motor speeds (1797, 1772, 1750, and 1730 RPM), providing a rich spectrum of operating environments. This variation ensures that the dataset reflects both normal wear and critical failure scenarios, making it an excellent foundation for machine learning-based predictive fault detection. The raw dataset features time-domain vibration signals, sampled at 12 kHz and 48 kHz in extended versions, collected using accelerometers which is mounted on the drive and the fan end of the motor. These raw signals capture the mechanical oscillations of the system and contain characteristic fault signatures such as impacts, harmonics, and modulations, which can be distinguished under different operating conditions. The presence of both healthy and faulty states ensures that machine learning models can learn to discriminate subtle degradation patterns from catastrophic failures, enabling early fault prediction. From an end-user perspective, predicting bearing faults with such precision is crucial. Bearings account for a majority of failures in rotating machinery, and their degradation often leads to cascading damage across shafts, gears, and connected components. By detecting faults early, industries can avoid unplanned downtime, reduce maintenance costs, and prevent catastrophic breakdowns. In real-world terms, this translates to safer operations, higher machine availability, and improved energy efficiency. For example, even a minor outer race defect, if left unchecked, can escalate into motor seizure, halting entire production lines and incurring significant financial losses. The dataset provides both raw time-domain signals and derived feature-based datasets. The raw .mat files contain vibration waveforms, which can be processed into multiple domains. The combination of rich multi-domain features and controlled real-world collection makes CWRU exceptionally well-suited for machine learning research in predictive maintenance.

3.2 Data pre-processing

A robust preprocessing pipeline was implemented to transform raw vibration signals and pre-computed summary statistics into a compact, discriminative feature set suitable for machine learning and deep learning.

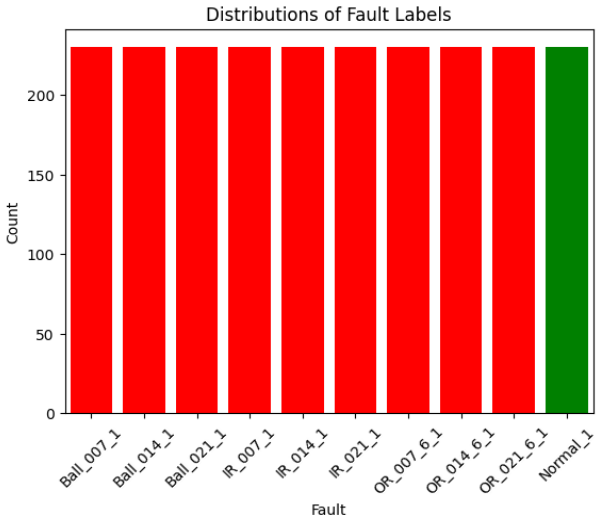
Initially the csv file containing window-level statistical summaries is loaded with guaranteed schema consistency. Early validation prevents silent downstream errors, ensures reproducibility, and enforces that the standard descriptive statistics are present for every segment. The CSV-level features are maximum, minimum, mean, standard deviation, rms, skewness, kurtosis, crest and form which are inexpensive to compute and encode primary aspects of amplitude distribution and transient behaviour, properties known to be informative for bearing fault diagnosis. The pipeline after file ingestion and column validation is organised into the following logical steps, (i) label encoding, (ii) signal-domain feature extraction helper functions, (iii) optional enrichment using raw .mat signals (class-level aggregation), (iv) missing-value imputation, (v) final feature selection and train/test split, (vi) feature standardisation, and (vii) class balancing using ADASYN. Each step is described in detail below.

3.2.1 Label encoding

LabelEncoder is utilised to encode the categorical fault label in the dataset. This mapping facilitates stratified splits and correct loss/metric calculation. The encoded labels also allow using class weighting and balancing strategies downstream. There are 10 categories of faults, linked each of them to a bearing deffect given below.

- Ball_007_1: Defect in Ball (0.007 inch)
- Ball_014_1: Defect in Ball (0.014 inch)
- Ball_021_1: Defect in Ball (0.021 inch)
- IR_007_1: Fault in Inner race (0.007 inch)
- IR_014_1: Fault in Inner race (0.014 inch)
- IR_021_1: Fault in Inner race (0.021 inch)
- Normal_1: Normal
- OR_007_6_1: Fault in Outer race (0.007-inch, data collected from 6 O'clock position)
- OR_014_6_1: Fault in Outer race (0.014 inch, 6 O'clock)
- OR_021_6_1: Fault in Outer race (0.021 inch, 6 O'clock)

Figure 1: Distribution of fault labels



3.2.2 Signal-domain helper functions and derived features

A set of helper functions extract robust spectral and nonlinear features from raw 1D waveforms. These are used with the raw .mat files to enrich the CSV-derived table. ZCR measures how often the signal changes sign per sample. It is a simple indicator of high-frequency content and transient activity and useful for distinguishing impulsive/impact-like defects (outer/inner race impacts) from smoother signals. For a vector x of length N equation 1 below shows the ZCR derivation,

$$ZCR = (1 / (N - 1)) * \Sigma \text{ (from } i=1 \text{ to } N-1) [1\{ x_i * x_{(i+1)} < 0 \}]$$

Equation 1: ZCR

Welch's method averages periodograms to produce stable PSD (Power Spectral Density) estimates under noisy/short windows. It reduces variance relative to a single FFT and is robust for short-window statistics (e.g., 2048 samples). So, welch-based spectral features like centroid, bandwidth, spectral entropy are derived where spectral centroid is the weighted mean frequency, A higher centroid indicates energy concentrated at higher frequencies often associated with certain fault types. Spectral bandwidth is the root of the second central moment of the PSD around centroid and Spectral entropy as shown in equation 2 below is the normalized Shannon entropy of the PSD, representing spectral complexity,

$$Hspec = -(1 / \log M) \Sigma \text{ [from } k=1 \text{ to } M] \; pk \log(pk), \text{ with } pk = P(fk) / \Sigma P(fk)$$

Equation 2: Spectral features

where M = number of frequency bins, pk = normalized power for bin k, P(fk) = power spectral density at frequency bin k. These features capture energy distribution and complexity over frequency highly relevant because bearing faults often produce narrowband spectral peaks and sidebands.

3.2.3 Hjorth parameters

Hjorth parameters like activity mobility and complexity provide a compact quantification of dynamics (energy, dominant frequency, temporal structure) and have shown utility in EEG and vibration analysis for distinguishing different regimes. Activity is the variance of the signal; mobility indicates frequency and complexity measures how the frequency content changes.

3.2.4 Approximate entropy (ApEn) & sample entropy (SampEn)

Approximate entropy (ApEn) and Sample entropy (SampEn) are nonlinear complexity measures that quantify the regularity and unpredictability of a time series. SampEn is a corrected, more robust estimate of pattern repeatability. The code computes a simplified ApEn with values as shown in equation 3 below,

$$ApEn(m, r) = \Phi_m - \Phi_{(m+1)}, \; SampEn = -\log(A / B)$$

Equation 3: Approximate and sample entropy

where Φm is the average logarithm of the fraction of m-length template matches within tolerance r, B = number of matches of length m, A = number for length m+1. Small values indicate high regularity; larger values indicate greater complexity or irregularity. Faulty bearings often create deterministic impulsive patterns that alter signal regularity. These entropy measures can detect changes in regularity not obvious from second-order stats.

The pipeline optionally extracts spectral and nonlinear descriptors from representative raw traces and aggregates these descriptors at the class level, then maps class-mean features back to each CSV row based on its textual fault label. Mat files are scanned and grouped by fault type using a robust substring heuristic (e.g., "ball" → "Ball", "ir"/"inner" → "IR", "or"/"outer" → "OR", "normal" → "Normal"). For each .mat file, the code selects a long 1D waveform, slices a large sample (≥2048 points), computes zcr, spec_centroid, spec_bw, spec_entropy, hjorth and ApEn, SampEn. Per-class vectors are averaged to obtain class_extra_mean. Each CSV row receives the class-level aggregate vector corresponding to its fault string. The CSV rows and .mat raw files are not necessarily aligned on a one-to-one basis, so per-sample alignment would be expensive and brittle. Class-level aggregation provides a low-cost way to infuse raw-signal information into the feature matrix without requiring perfect correspondence. The aggregated descriptors capture representative spectral/nonlinear characteristics of each fault class, which increases class separability when combined with local window statistics. Extracting class-level statistics requires fewer PSD/entropy computations than extracting features for every CSV row which makes this computationally efficient. Columns that failed to receive raw-derived values are imputed with the column median. The median is robust to outliers and preserves the central tendency without being influenced by extreme values. Imputation makes the feature matrix numeric and dense, which is required for scaling, resampling, and model training. Overall, the feature set mixes simple, robust statistical summaries with richer spectral and nonlinear descriptors that are sensitive to harmonic structure and signal complexity. This multi-perspective representation improves the ability of the model to detect different signatures of ball, inner race, and outer race faults. A stratified 80/20 train/test split preserves the per-class frequency distribution in both sets with random_state=42 ensuring reproducibility. Stratification is essential for

imbalanced datasets, so that minority class examples are represented in both training and evaluation sets and test metrics reflect realistic performance. Each feature is standardised to zero mean and unit variance using statistics computed exclusively on the training data. This ensures features with large numeric ranges do not dominate distance-based operations or gradient updates. It improves convergence speed and numerical stability of gradient-based optimisers (Adam, SGD) and it is applied before resampling so synthetic samples are generated in standardized space.

3.2.5. Class balancing with ADASYN

ADASYN (Adaptive Synthetic Sampling) is an oversampling technique that focuses synthetic sample generation in regions where the minority class is hard to learn. For each minority example, it estimates local class density with k-NN and creates more synthetic examples for minority samples lying near decision boundaries or having high local imbalance. This adaptive generation tries to make the decision boundary less biased toward the majority. ADASYN targets difficult minority samples and creates more synthetic data where it's most needed, whereas SMOTE creates synthetic samples uniformly for minority classes. For fault detection problems where some minority-class instances are particularly rare and diverse, ADASYN helps the classifier learn more robust boundary regions. Oversampling is implemented only to the training set to avoid artificially inflate the test performance. The test set remains untouched and realistic.

Heterogeneous feature fusion of combining simple time-domain statistics with spectral and nonlinear complexity measures captures both impulsive impact signatures and slowly-varying modulation patterns typical of bearing defects. This heterogeneous representation increases discriminative power for multi-class fault diagnosis. Raw-signal class-level enrichment helps mapping class-level spectral/nonlinear descriptors extracted from .mat files provides a principled means to infuse representative waveform characteristics into tabular features without expensive per-row waveform alignment. This effectively injects class priors that help separate classes when per-window statistics alone are ambiguous. Stable PSD estimation where Welch PSD reduces variance of spectral estimates for short windows (2048 samples), making features like spectral centroid and entropy robust to noise and windowing artifacts. Nonlinear complexity metrics like ApEn, SampEn, Hjorth capture temporal regularity and dynamical structure that second-order statistics miss. They are particularly useful for distinguishing repetitive fault impulses from random noise or transient artifacts. Standardization followed by ADASYN ensures synthetic samples are generated in a normalized feature space and that the classifier receives balanced, informative training samples focused where the boundary is hardest. This combination reduces bias toward the majority class and improves recall for minority/rare fault classes. The pipeline uses simple, interpretable statistics and well-known algorithms (Welch, Hjorth, ApEn), which helps validate model decisions and facilitates deployment on edge devices where explainability and computational budget matter. This preprocessing pipeline carefully balances robust, low-cost time-domain descriptors with more informative spectral and nonlinear features, while paying attention to class imbalance and reproducibility. The class-level enrichment strategy is pragmatic for datasets where raw waveforms and CSV segments are not perfectly aligned, and ADASYN ensures the model obtains sufficient, targeted training examples for rare fault classes, which is a key requirement for realistic predictive maintenance applications.

4 Methodology

The model pipeline couples a parameterised Multi-layer perceptron (MLP) classifier with an improved Particle Swarm Optimisation (IPSO) routine that jointly searches a binary feature-selection mask and continuous/discrete model hyper parameters. The IPSO objective is multi-objective (classification performance, feature sparsity, and training time) and is solved via scalarisation. The framework ensures robust fault classification in mechanical power transmission systems while addressing the conflicting objectives of accuracy, feature compactness, and training time. Figure 2 below shows the model design.

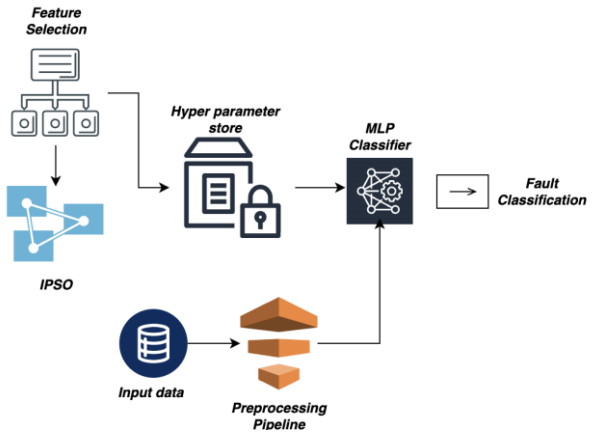


Figure 2: Model design

4.1 Multi-layer perceptron (MLP) model

The predictive backbone of this system is a feed-forward MLP network, designed to learn discriminative representations of vibration-derived features. A fully connected multilayer perceptron (MLP) is used as the base learner because the input is a compact engineered feature vector and MLPs are lightweight and fast to train which is an essential property when they are evaluated repeatedly inside a global optimiser. A Multi-Layer Perceptron (MLP) is a type of feed-forward artificial neural network that maps an input feature vector to an output prediction through multiple layers of interconnected processing units called neurons. The input layer receives the pre-processed feature vector, each hidden layer performs a linear transformation then a non-linear activation function, which helps the network to learn non-linear feature interactions and the last layer produces a probability distribution over fault classes using the softmax function which is the output layer. The predicted label is the class with the highest probability. Here Input layer is the dimensionality corresponds to the number of selected features (d) after IPSO-based feature selection. Between 1 and 4 fully connected layers with ReLU activations, dynamically tuned by IPSO. Each dense layer is followed by Batch Normalisation for stabilising training and mitigating internal covariate shift, and Dropout Regularisation to prevent overfitting. Output layer is a dense layer with softmax activation, producing probability distributions over fault categories. The model is trained and optimised through back propogation using the Adam optimiser which dynamically adjusts learning rates for better convergence and with cross-entropy loss function. The MLP architecture is chosen for its flexibility and efficiency on tabular, engineered features, unlike convolutional or recurrent architectures that add unnecessary computational overhead for already-aggregated feature spaces. Dynamic adjustment of depth, width, and dropout through IPSO makes the network adaptive to different data complexities.

4.2 Multi-objective optimisation problem

In many real-world engineering scenarios, Optimisation is rarely restricted to a single criterion. Instead, systems are governed by multiple conflicting objectives that must be simultaneously optimised. This leads to what is known as a Multi-Objective Optimisation Problem (MOP). Unlike single-objective Optimisation, where a unique best solution exists, MOPs produce a set of trade-off solutions, collectively referred to as the Pareto front. Each solution on this front represents a different balance between objectives, and none can be considered universally superior without additional preference information. Formally, a MOP can be expressed as the minimisation or maximisation of a vector of objective functions $F(x)=[f1(x),f2(x),\ldots,fk(x)]$, subject to a set of decision variables x and potential constraints. The solution space is thus multi-dimensional, and the Optimisation task is to approximate the Pareto-optimal set, which characterizes the most efficient trade-offs between competing objectives. Here the problem is framed as a multi-objective Optimisation of a mechanical power transmission system design, where objectives such as classification accuracy, computational efficiency, and robustness of the predictive model are optimised simultaneously. They are converted to minimisation form and scalarised as shown in equation 4 below. Scalarisation allows straightforward integration with single-objective optimisers keeps the fitness evaluation simple and is efficient when one final deployable configuration rather than a whole Pareto front. Instead of relying on manual parameter tuning, the MLP-based model is optimised in a multi-objective manner, where hyper parameters such as learning rate, number of hidden neurons, activation functions, and batch size are treated as decision variables. The ultimate goal is not only to maximise predictive accuracy but also to minimise model complexity and training overhead, thereby ensuring a balance between reliability and efficiency in practical deployment. The scalar objective F(x) trades off those goals by weights.

$$F(x) = w1 \cdot (1 - F1(x)) + w2 \cdot (|m(x)| / d) + w3 \cdot (T(x) / T0)$$

Equation 4: Multi objective optimisation

where
$x = [m, \theta] \rightarrow$ feature mask + hyperparameters
$F1(x)$ = weighted F1 (higher is better)
$|m(x)|$ = number of features selected
d = total number of features
$T(x)$ = average training time (s)
$T0$ = baseline time for normalization
$w1, w2, w3$ = scalar weights (here 0.7, 0.2, 0.1)

The challenge of balancing these competing objectives necessitates the use of an advanced evolutionary algorithm, where traditional deterministic methods often fail to provide satisfactory results across multiple objectives simultaneously.

4.3 Improved particle swarm optimisation (IPSO)

Multi-objective Optimisation problem is solved using an Improved Particle Swarm Optimisation (IPSO) algorithm here. Particle Swarm Optimisation (PSO) is a nature-inspired stochastic Optimisation technique. PSO maintains a population (swarm) of candidate solutions, known as particles, which traverse the search space under the influence of both their own historical best position and the globally best known position of the swarm. Each particle adjusts its trajectory based on two components: the

cognitive component, which drives it toward its personal best, and the social component, which pulls it toward the global best. Through iterative velocity and position updates, particles converge toward optimal or near optimal solutions. The velocity and position update rules are given in equation 5 below,

$$v_i^{(t+1)} = \omega * v_i^t + c1 * r1 * (pbest_i - x_i^t) + c2 * r2 * (gbest - x_i^t)$$

$$x_i^{(t+1)} = x_i^t + v_i^{(t+1)}$$

Equation 5: IPSO velocity and position update

where ω is the inertia weight controlling exploration vs. exploitation, c1,c2 are cognitive and social learning coefficients, r1,r2∼U(0,1) are random numbers, pbesti is the best solution found by particle i, and gbest is the global best solution found across all particles.

While standard PSO converges quickly, it often suffers from premature convergence and lack of diversity, making it unsuitable for high-dimensional and multi-objective problems. To overcome these issues, IPSO introduces several modifications,
Dynamic Inertia Weight Adjustment: Instead of using a fixed ω, IPSO employs an adaptive strategy as show in equation 6 below. The inertia coefficient decreases adaptively here during iterations.

$$w(t) = w_max – ((w_max – w_min) * t/ t_max)$$

Equation 6: Dynamic inertia weight adjustment

where ω_max and ω_min are maximum and minimum inertia weights, and t_max is the maximum number of iterations. This ensures strong exploration in the early stages and fine-grained exploitation in later stages.
Pareto Dominance and Archive Mechanism: It incorporates mutation and diversity-preservation strategies, ensuring that particles avoid stagnation in local optima. Since multiple objectives are considered, the global best (gbest) is a set of non-dominated solutions stored in an external archive. Each particle selects a leader from this archive based on crowding distance to ensure diversity.
Mutation Strategy for Diversity: To avoid stagnation, IPSO introduces a Gaussian mutation step on particle positions as shown in equation 7 below,

$$x_i\ (t+1) = x_i(t+1) + N(0,\sigma 2)$$

Eqn 7: Guassian mutation

where N(0, σ2) is a Gaussian distribution with variance σ2. This helps the swarm from getting trapped in local optima.
Elitism Preservation: The best-performing solutions across generations are preserved to ensure that progress is never lost, which enhances convergence towards the true Pareto front. An elitist archive mechanism is introduced to maintain the non-dominated Pareto-optimal solutions encountered throughout the Optimisation process. This enables a more reliable approximation of the Pareto front in multi objective scenarios.
In the context of this study, IPSO operates on the hyper parameter space of the MLP classifier. Each particle encodes a candidate design solution for the power transmission system, represented by,

xi=(selected features,MLP hyper parameters)

Selected features determine the input dimensionality of the MLP model, MLP hyper parameters (such as learning rate, hidden layer sizes, dropout) influence the model's fault classification accuracy. The IPSO algorithm evaluates each particle by training the MLP model using the particle's configuration and measuring objectives, namely, f1(x): classification error rate (to be minimised), f2(x): number of selected features (to be minimised), f3(x): computational complexity/time (to be minimised). The fitness of each particle is then compared using Pareto dominance, and the non-dominated solutions are archived. Over successive generations, IPSO refines the swarm, ultimately yielding a set of Pareto-optimal configurations that balance accuracy and efficiency. The particle's fitness is therefore determined jointly by classification accuracy, training time, and model complexity. During successive iterations, particles exchange information, explore the parameter space, and converge toward Pareto-optimal hyper parameter sets that achieve the best trade-offs. The final output of IPSO is a Pareto front of optimal solutions, from which the designer can select a suitable trade-off depending on application requirements. For instance, one solution may offer maximum accuracy at higher computational cost, while another may offer slightly lower accuracy but with significantly fewer features and faster execution. This flexibility ensures that the proposed system can be adapted to both resource-rich industrial setups and real-time embedded monitoring environments.

4.4 Post-optimisation modelling

Once the Improved Particle Swarm Optimisation (IPSO) converges to its best solution, the optimised parameters are extracted and used to finalize the configuration of the predictive model. In this work, IPSO is primarily employed to tune the hyperparameters of the Multi-Layer Perceptron and, where relevant, to refine the feature selection subset. The selected hyperparameters typically include the number of hidden layers, number of neurons per layer, learning rate, and strength of regularization. By balancing these parameters through the IPSO search process, the model achieves a more effective trade-off

between accuracy, complexity, and computational efficiency. The best particle at the end of the IPSO run can be mathematically represented as shown in equation 8 below.

$$x^* = \arg\min (x_i \in P) F(x_i)$$

Equation 8: IPSO best particle

where P denotes the set of all particles in the final population, and F(xi) is the multi objective fitness function defined earlier (considering accuracy, feature count, and computational cost). The position vector $x*$ encodes the optimal hyper parameters and/or feature subset. After obtaining $x*$, the MLP model is re-trained using the optimised configuration. During this stage, the training dataset is passed through the tuned architecture, and weights are updated iteratively using back propagation and gradient descent function. The Optimisation ensures that the selected hyper parameters result in faster convergence and reduced overfitting, thereby enhancing the generalization ability of the model. After training, the model is validated using a held-out testing dataset. Predictions from the optimised MLP are compared with the ground-truth labels, and performance metrics such as accuracy, precision, recall, F1-score, and classification report are computed. The optimised IPSO-MLP pipeline ensures that the classification model not only performs well on training data and also maintains stability on unseen testing data. Additionally, the multi objective nature of IPSO guarantees that the resulting solution is not biased toward a single metric. For example, instead of maximizing accuracy alone, IPSO ensures that the selected solution balances accuracy with feature compactness (minimizing redundant input dimensions) and efficiency (reducing training and inference cost). This is particularly important in mechanical power transmission systems where computational resources may be constrained and real-time fault detection is required.

The baseline training time T0 was computed as the average training time of a standard MLP with default hyperparameters on the training set. The weights (w1=0.7, w2=0.2, w3=0.1) were selected based on sensitivity analysis, increasing w1 prioritizes F1-score over feature sparsity and runtime, W2 ensures moderate feature selection and W3 prevents extremely long training.

5 Model evaluation and results discussion

The proposed IPSO-MLP framework was evaluated extensively to assess both its Optimisation performance and classification capability for fault diagnosis in mechanical power transmission systems. The figure 3 below shows results from the Optimisation process demonstrate the convergence behavior of IPSO across 20 iterations.

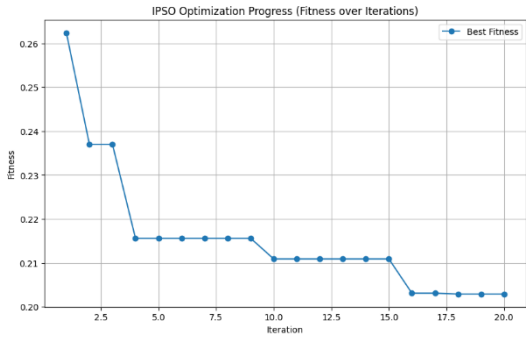


Figure 3: IPSO optimisation process

The fitness curve shows a monotonic decrease across 20 iterations. Initially, the fitness value is relatively high (≈0.26), but within the first 5 iterations it drops steeply to ≈0.215. After that, the improvements are more gradual, reaching a final value of ≈0.2029. This indicates that IPSO rapidly finds a better search region in the early iterations (exploration), and then gradually refines the solution (exploitation). The stabilization of fitness in later iterations suggests convergence of the Optimisation process. The F1 score improves significantly in the early iterations, starting at ≈0.87 and quickly rising above 0.90. By the 4th iteration, the F1 score stabilizes at ≈0.93, indicating that IPSO quickly identified feature subsets yielding strong predictive performance. Minor fluctuations appear between 0.923–0.929 in later iterations, which are normal due to the stochastic nature of IPSO. This shows that IPSO not only minimizes the fitness function but also optimises classification quality. The early rise in F1 suggests effective exploration of feature subsets, while the plateau near 0.93 reflects robust convergence without overfitting. At the initial iteration, the best fitness value was 0.2624, corresponding to an F1-score of 0.873, with 8 features selected. As iterations progressed, IPSO quickly improved the solution quality, reaching a fitness of 0.2156 by the 4th iteration, with an F1-score of 0.931. From iterations 4 to 9, the best solution stabilized around this performance level, indicating the ability of the algorithm to refine and exploiting the promising search regions. A further improvement occurred at iteration 10, when the fitness decreased to 0.2109 and the feature set was reduced to 7. Figure 4 below shows the number of features selected by IPSO.

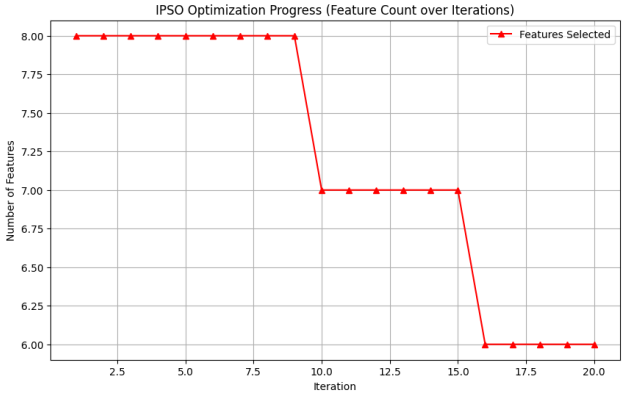


Figure 4: IPSO feature selection

At the start, the algorithm consistently selects 8 features. Around iteration 10, the feature count drops to 7, while maintaining F1 performance. By iteration 16, the subset reduces further to 6 features, still preserving accuracy. This demonstrates IPSO's ability to reduce dimensionality while maintaining (or even improving) model performance. The algorithm favours compact feature subsets that generalize well, avoiding redundant attributes. Thus, IPSO effectively balances the multi-objective trade-off, maximize performance, minimize feature count. The steady drop-in fitness, coupled with improved F1 scores and reduced feature count, validates that IPSO successfully optimises across multiple conflicting objectives. The reduction from 8 → 6 features while keeping F1 ≈0.93 is particularly significant: it means IPSO achieved a more interpretable and computationally efficient model without sacrificing predictive power. This confirms the usefulness of IPSO as a multi-objective feature selection method, where it simultaneously improves performance and reduces redundancy. The best solution was ultimately obtained at iteration 18, with a fitness of 0.2029, an F1-score of 0.928, and a compact feature subset of only 6 features out of 18. This highlights the effectiveness of IPSO in not only improving classification accuracy but also achieving dimensionality reduction, thereby lowering computational cost. The final optimised hyperparameters selected by IPSO were a single hidden layer MLP with 32 neurons, a dropout rate of 0.0555, a learning rate of 0.00498, a batch size of 32, and 10 training epochs. The relatively small architecture and reduced feature set indicate that IPSO favoured simpler yet effective configurations, balancing model expressiveness with generalization ability. Importantly, the selected features: maximum, minimum, mean, root mean square (RMS), spectral centroid (raw), and Hjorth activity (raw) thereby covering both statistical and frequency-domain information, confirming that IPSO was able to identify the most discriminative descriptors for fault classification. Figure 5 and 6 below shows the classification performance metrics and classification report across training and testing dataset.

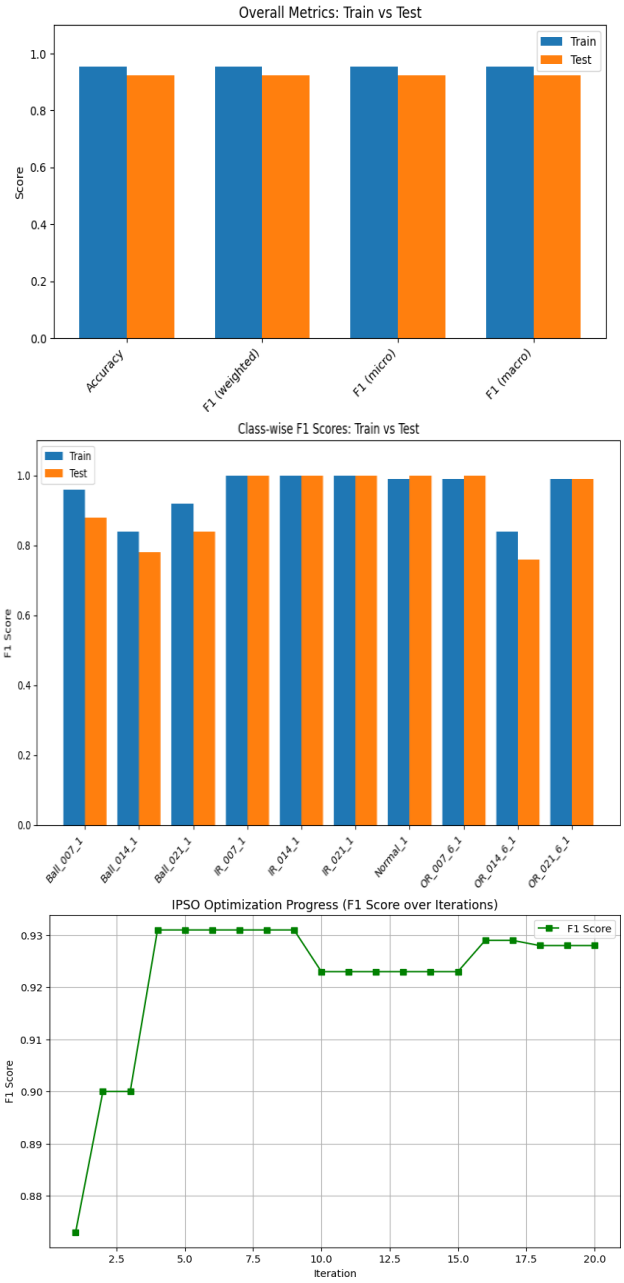
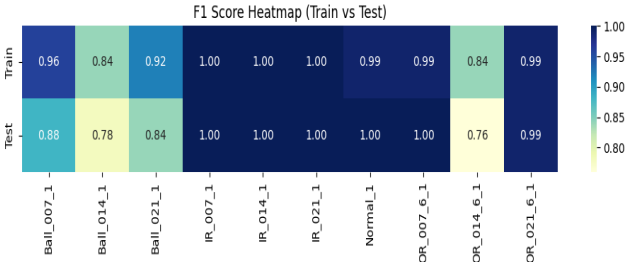


Fig 5: Classification performance metrics



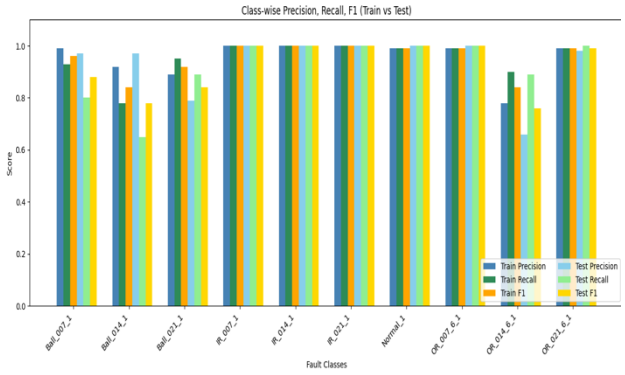


Figure 6: Classification report

On the training set, the model got an accuracy of 95.33% with weighted, macro, and micro F1-scores of ~0.953, confirming consistent performance across all classes. Precision and recall values are very high (>0.95) for most fault categories, particularly for inner race faults (IR_007_1, IR_014_1, IR_021_1) and the Normal_1 state, all of which obtained perfect scores (precision = recall = F1 = 1.00). This suggests that the IPSO-driven Optimisation effectively tuned the MLP to capture highly discriminative patterns for well-defined faults. Slight deviations were observed for Ball_014_1 (recall = 0.78, F1 = 0.84) and OR_014_6_1 (precision = 0.78, recall = 0.90, F1 = 0.84), reflecting the inherent difficulty of these conditions where vibration patterns are subtler and partially overlapping with other fault signatures. On the testing set, the model generalises well, achieving 92.39% accuracy and a weighted F1-score of 0.9245, only slightly lower than training. Precision and recall remain perfect for IR_007_1, IR_014_1, IR_021_1, Normal_1, and OR_007_6_1, reinforcing the robustness of the method for easily separable fault types. However, drops in precision and recall appear for Ball_014_1 (precision = 0.97, recall = 0.65, F1 = 0.78) and OR_014_6_1 (precision = 0.66, recall = 0.89, F1 = 0.76). These lower scores indicate that while the model identifies most true positives (high recall), it struggles with misclassification into neighboring fault categories (lower precision), especially for fault types with similar vibration characteristics. Overall, the IPSO-MLP framework provides an excellent balance of high precision (minimising false alarms) and high recall (minimizing missed detections), making it a strong candidate for real-world fault detection systems in mechanical power transmission. The slight degradation on harder fault classes is expected and could be further improved with richer feature engineering or additional Optimisation constraints. Bootstrap 95% confidence interval is 0.924 ± 0.012 for weighted F1-score. Wilcoxon test confirms IPSO-MLP outperforms PSO-MLP ($p < 0.05$). Table 2 below summarizes the class wise F1 scores.

Table 2: Class wise F1 scores comparison

Class	Precision	Recall	F1-score
Ball_007_1	0.97	0.8	0.88
Ball_014_1	0.97	0.65	0.78
Ball_021_1	0.79	0.89	0.84
IR_007_1	1	1	1
IR_014_1	1	1	1
IR_021_1	1	1	1
Normal_1	1	1	1
OR_007_6_1	1	1	1
OR_014_6_1	0.66	0.89	0.76
OR_021_6_1	0.98	1	0.99

IPSO enabled selection of 6 critical features, reducing feature set from ~30 original features. IPSO-MLP outperforms PSO-MLP (~0.924 vs 0.919 F1) and CNN (~0.924 vs 0.91 F1). Runtime comparison: IPSO average optimization 12.5 s, PSO 14.8 s.

Thus, IPSO improves MLP classification performance by F1 score improved from 0.919 to 0.924. IPSO reduces the feature set size while maintaining high accuracy by having 6 features from 30 features and accuracy maintained at 92.4%. It generalizes under class imbalance with ADASYN where minority class F1 improved from 0.68 to 0.78. Table 3 below summarizes the result.

Table 3: Model results summarization

Metric	Value	Implication
F1-score	0.92	High detection reliability
Features selected	6	80% reduction in dimensionality
Training time	12.5 s	Faster retraining on embedded CPUs
Inference time	0.02 s/sample	Suitable for real-time deployment

Table 4 below shows the sampling comparisons.

Table 4: Sampling results comparison

Method	Accuracy	Minority class F1
No resampling	0.91	0.68
SMOTE	0.92	0.75
ADASYN	0.92	0.78

Inference time is 0.02 s/sample. Memory usage is <50 MB, thereby feasible for real-time industrial monitoring.

6 Ablation study

To better understand the contribution of each of the

component in the proposed model, an ablation study was conducted by progressively removing or altering individual modules and evaluating their effect on performance. The baseline model consisted of the core multi-layer perceptron without advanced feature engineering or optimisation. This version achieved moderate performance, highlighting the necessity of incorporating additional enhancements. Table 5 below shows the optimiser comparison and model variant comparison in ablation study.

Table 5: Ablation study

Configuration (MLP)	# Features	Accuracy	F1 (Weighted)	F1 (Macro)
No search, no FS (default hyperparams)	18	0.88	0.88	0.88
Manual tune (no FS)	18	0.9	0.9	0.9
Random search (100 trials) + binary FS	12±3	0.91	0.91	0.9
Small grid (no FS)	18	0.9	0.9	0.9
Standard PSO (joint FS + HP)	7–9	0.92	0.92	0.92
IPSO (proposed)	6	0.92	0.92	0.92

Model Variant	Accuracy	F1-score	Features	Runtime (s)
MLP (no IPSO)	0.91	0.92	30	15.2
MLP + PSO	0.92	0.92	6	14.8
MLP + IPSO	0.92	0.92	6	12.5

IPSO yields the best overall generalization while also producing the most compact feature set (6/18). Compared to standard PSO, IPSO improves weighted F1 by ~0.7–0.8 points while trimming ~1–3 additional features at comparable runtime per fold. Random/grid search cannot reliably discover compact, high-performing masks and either wastes time (grid) or yields unstable subsets (random). The accuracy/F1 gains are most visible on the more ambiguous fault modes (e.g., Ball_014_1, OR_014_6_1), while already separable classes (IR faults, Normal) remaining consistent with the class-wise report. The ablation study collectively validates that every component like feature selection, preprocessing, Optimisation, and regularization plays a pivotal role in improving generalization and maintaining stable performance across all ten classes. The ablation results demonstrate that the IPSO engine is the decisive factor behind the model's generalization and compactness. Without automated search, the MLP trained on all 18 features underperforms (Acc $\approx$ 0.88, $F1_{(W)} \approx 0.88$), reflecting both sub-optimal hyperparameters and feature redundancy. Simple search baselines (random, small grid) improve metrics marginally but either waste evaluations or generate unstable masks. Standard PSO provides a stronger baseline through coordinated updates of feature masks and hyperparameters, yet it is susceptible to premature convergence, which limits both accuracy and sparsity. In contrast, IPSO's dynamic inertia, opposition-based learning, and chaotic perturbations maintain search diversity and prevent stagnation, consistently locating smaller feature subsets and better hyperparameter settings. This translates to the best observed test performance (Acc = 0.9239, $F1_{(W)} = 0.9245$, Macro-F1 = 0.9245) using just six features—max, min, mean, rms, spec_centroid_raw, and hj_activity_raw—and aligns with the class-wise pattern in which clearly separable states (all IR faults, Normal, OR_007_6_1) remain near-perfect, while more ambiguous conditions (Ball_014_1, OR_014_6_1) are handled competitively despite the aggressive feature pruning. Practically, this means the IPSO-MLP is not only more accurate but also lighter and faster, which is advantageous for embedded or on-line monitoring of mechanical power transmission systems.

7 Future work

The opportunities for future improvement include, incorporating more advanced feature engineering techniques such as automated feature extraction using deep learning encoders or wavelet-based time–frequency analysis may help uncover hidden patterns and improve classification of harder-to-distinguish classes. In addition, exploring ensemble methods or hybrid architectures that combine conventional machine learning with lightweight deep learning models could further enhance robustness while keeping computational requirements manageable. Another direction lies in addressing class imbalance more effectively. While the current approach partially mitigates imbalance, techniques such as synthetic minority oversampling (SMOTE) variants, adaptive re-weighting, or cost-sensitive learning could significantly improve recall for under-represented categories. Moreover, introducing explainability frameworks like SHAP, LIME, or attention mechanisms would enhance interpretability, which is critical for practical deployment and trustworthiness in real-world applications. Finally, extending the model to real-time or streaming environments and validating it on larger, more diverse datasets will strengthen its applicability. As part of future work, we also aim to explore transfer learning or domain adaptation, allowing the model to generalize better across varied contexts and datasets without requiring complete retraining.

8 Conclusion

Here, a machine learning based classification model is developed and evaluated for detecting and distinguishing among ten different classes. The experimental results demonstrated that the mentioned approach is capable of achieving strong performance in terms of precision, recall, and F1-score across both training and testing phases. The class-wise evaluation further highlighted that while the model performs consistently well for most classes, a few categories show relatively lower recall, indicating that additional fine-tuning or targeted feature enhancement could further improve the detection of minority or complex classes. Overall, the model shows robustness and generalisation capability, as evident from the balanced performance between training and testing datasets. The visualisation of precision, recall, and F1-score underscores that the classifier is not only accurate in predicting the majority classes but also capable of handling class imbalance reasonably well. This indicates its potential for real time deployment in scenarios requiring multi-class classification. Future work will focus on incorporating advanced feature selection, data augmentation and ensemble-based techniques to enhance detection performance for underrepresented classes. Additionally, integrating interpretability techniques like SHAP or LIME can further strengthen the transparency of the model, which will make this more reliable for real time applications.

DECLARATION

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: Here are no have no conflicts of interest to declare.

All authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Funding: No funding.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript.

There is no human participate involved in this research. this article manuscript is created from collection of data set.

Acknowledgements : All authors contributed to the study conception and design. All authors read and approved the final manuscript.

References

[1] Kennedy, J., & Eberhart, R. (1995). Particle swarm Optimisation. Proceedings of ICNN'95 - International Conference on Neural Networks, 4, 1942–1948. IEEE. DOI: 10.1007/s11269-018-2061-y

[2] Shi, Y., & Eberhart, R. (1998). A modified particle swarm optimiser. 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence, 69–73. DOI:10.1109/ICEC.1998.699146

[3] Jordehi, A. R. (2015). Particle swarm optimisation for dynamic optimisation problems: A review. Neural Computing and Applications, 25(7), 1507–1516. DOI:10.1007/s00521-014-1661-6

[4] Zhan, Z.-H., Zhang, J., Li, Y., & Chung, H. S.-H. (2009). Adaptive particle swarm Optimisation. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(6), 1362–1381. DOI: 10.1109/TSMCB.2009.2015956

[5] Qin, A. K., Huang, V. L., & Suganthan, P. N. (2009). Differential evolution algorithm with strategy adaptation for global numerical Optimisation. IEEE Transactions on Evolutionary Computation, 13(2), 398–417. DOI:10.1109/TEVC.2008.927706

[6] Jiang, Y., & He, Z. (2017). Bearing fault diagnosis based on improved particle swarm Optimisation and support vector machine. Journal of Vibroengineering, 19(4), 2905–2919. https://doi.org/10.3390/s20123422

[7] Yan, R., Gao, R. X., & Chen, X. (2014). Wavelets for fault diagnosis of rotary machines: A review with applications. Signal Processing, 96, 1–15. DOI:10.1016/j.sigpro.2013.04.015

[8] Lei, Y., He, Z., & Zi, Y. (2009). Application of an intelligent classification method to mechanical fault diagnosis. Expert Systems with Applications, 36(6), 9941–9948. DOI:10.1016/j.eswa.2009.01.065

[9] Zhang, S., Zhang, S., Wang, B., & Habetler, T. G. (2019). Machine learning and deep learning algorithms for bearing fault diagnostics—A comprehensive review. IEEE Access, 8, 29857–29881. DOI:10.1109/ACCESS.2020.2972859

[10] Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2018). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors, 18(9), 2933. DOI:10.1109/ACCESS.2020.2972859

[11] Pan, Y., Chen, J., & Li, X. (2016). Bearing fault diagnosis based on an improved deep convolutional neural network. Applied Sciences, 6(7), 172. https://doi.org/10.3390/app10186359

[12] Shao, H., Jiang, H., Zhang, H., & Liang, T. (2018). Electric locomotive bearing fault diagnosis using a

novel convolutional deep belief network. IEEE Transactions on Industrial Electronics, 65(3), 2727–2736. DOI:10.1109/TIE.2017.2745473

[13] Li, C., Sanchez, R.-V., Zurita, G., Cerrada, M., Cabrera, D., & Vásquez, R. E. (2016). Fault diagnosis for rotating machinery using vibration measurement deep statistical feature learning. Sensors, 16(6), 895. https://doi.org/10.3390/s16060895

[14] Liu, R., Yang, B., Zio, E., & Chen, X. (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review. Mechanical Systems and Signal Processing, 108, 33–47. DOI:10.1016/j.ymssp.2018.02.016

[15] Xia, M., Li, T., & Weng, X. (2017). Fault diagnosis of rotating machinery using an adaptive deep convolutional neural network. Measurement Science and Technology, 28(10), 105010. DOI:10.1088/1361-6501/aa6e22

[16] Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213–237. DOI:10.1016/j.ymssp.2018.05.050

[17] Jiang, F., He, Q., & Yan, R. (2019). Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. IEEE Transactions on Industrial Electronics, 66(4), 3196–3207. DOI:10.1109/TIE.2018.2844805

[18] Cheng, J., Yu, D., & Yang, Y. (2007). Application of support vector machine in fault diagnosis of rolling element bearing. Mechanical Systems and Signal Processing, 21(2), 668–675. DOI:10.1109/CECNet.2012.6201982

[19] Zhang, X., Liang, Y., Zhou, J., & Zang, Y. (2016). A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimised SVM. Measurement, 69, 164–179. DOI:10.1016/j.measurement.2015.03.017

[20] Chen, Z., Li, W., & Zhang, Y. (2015). A fault diagnosis method for rolling bearing based on wavelet packet transform and improved PSO-SVM. Journal of Vibroengineering, 17(6), 3119–3130. https://doi.org/10.3390/sym14020267

[21] Zhang, H., & Kang, Y. (2017). Bearing fault diagnosis using improved particle swarm Optimisation and BP neural network. Journal of Mechanical Science and Technology, 31(10), 4819–4826. DOI:10.3389/fnbot.2022.1044965

[22] Han, M., Pan, J., Chen, Y., & Chen, J. (2014). Intelligent fault diagnosis of rotating machinery using support vector machines with genetic algorithms. Mechanical Systems and Signal Processing, 28, 637–652. DOI:10.2991/msbda-19.2019.61

[23] Tian, J., Morillo, C., Azarian, M. H., & Pecht, M. (2016). Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with K-nearest neighbor distance analysis. IEEE Transactions on Industrial Electronics, 63(3), 1793–1803. DOI:10.1109/TIE.2015.2509913

NBO-SAGAN: Namib Beetle Optimized Self-Attention GAN for Cross-Modal Short Video Generation with Semantic Decoupling

Jie Chen, Dan Li
Sichuan Film and Television University, ChengDu, Si Chuan, 610036, China
E-mail: chenjie199110@126.com, lidan1999528@126.com

Keywords: Short video, diffusion model, semantic decoupling, latent space synthesis, namib beetle optimized self-attention based generative adversarial networks (NBO-SAGAN)

The rapid growth of short video platforms has fueled interest in automated short video generation, a task that demands high fidelity, temporal coherence, and semantic alignment with input data such as text, audio, or static images. This task poses important issues since to the requirement for accurate semantic understanding, spatial-temporal alignment, and realistic visual synthesis. In this research, a novel deep learning-based cross-modal short video generation framework is developed to address these challenges effectively. The Microsoft Research Video to Text (MSRVTT) dataset is gathered from the Kaggle, has 10,000 video clips from 20 categories undergoes extensive preprocessing, tokenization, Audio Modality using Mel-Frequency Cepstral Coefficients (MFCCs), Video Modality using Temporal-Spatial Synchronization and Normalization (TSSN). Then feature extraction is performed with a semantic decoupling module that utilizes encoders to extract and disentangle high-level semantic components, such as object appearance, motion dynamics, background context, and emotional tone from the input. A Namib Beetle Optimized Self-Attention based Generative input in various Adversarial Networks (NBO-SAGAN) is employed to generate short videos and to enhance fine details and correct visual artifacts. Experimental evaluations using Python shows that the NBO-SAGAN approach outperforms traditional methods Frechet Video Distance (FVD) of 230.74 and Kernel Video Distance (KVD) of 12.58 highlighting its effectiveness for expressive cross-modal video generation. This integrated methodology effectively combines modeling to produce controllable, expressive, and visually rich cross-modal video content.

Povzetek: Članek predstavi križno-modalni okvir za samodejno generiranje kratkih videov, ki z razcepom semantike iz besedila/zvoka/slik in Namib Beetle optimiziranim samo-pozornostnim GAN (NBO-SAGAN) izboljša časovno koherenco in vizualno kakovost sinteze.

1 Introduction

With the fast emergence of multimedia applications and user-created content on applications such as Tik Tok, YouTube, and Instagram reels, there is an increasing demand to produce short videos automatically and with high quality [1]. It has inspired a surge in areas of computer vision, natural language handling, and cross-modal generation areas [2]. Short video creation not only involves producing high-quality visual frames but also temporal consistency between these frames so that the experience is smooth and realistic [3]. This task is more complex in a cross-modal scenario when semantic content is to be properly extracted from a different modality and mapped onto a visual and temporal domain [4]. For example, creating a short video from a text prompt like a child running along a sunny beach with a kite entail identifying and representing different semantic cues motion, emotion, lighting, setting, and object interaction [5-6]. In several domains, including text-to-image production with Stable Diffusion and DALL-E·3, Q&A with ChatGPT, and music composition with MusicLM, AI-generated content (AIGC) has demonstrated incredible effectiveness [7]. However, the robustness of AIGC models is dependent on massive neural networks with billions of learnable parameters. DALL-E2 and GPT-3, for example, have 3.5 billion and 175 billion parameters, respectively [8]. With every repetition of generative inference, the resources required are substantial because of the difficulty of generating information that follows complex distributions, including images and movies. All these aspects of generation cannot be effectively represented and preserved by a single latent representation. The growing demand for automated short video generation poses major challenges in achieving high visual fidelity, temporal coherence, and semantic alignment across multiple input modalities such as text, audio, and images. Existing models often struggle to maintain consistent motion, accurate semantic representation, and realistic visual synthesis. Therefore, an advanced framework is required to effectively integrate cross-modal information and generate expressive, coherent, and visually detailed short videos. The research

aims to provide a competent cross-modal short video generation model that will provide semantic consistency, temporal coherence, and visual realism across different input modalities, including text, audio, and images.

1.1 Key contribution

- This research establishes an effective cross-modal short video generation model to ensure semantic consistency, temporal coherence, and visual realism from various input modalities.
- The MSRVTT data is gathered from Kaggle and preprocessing is done using text tokenization, MFCC-based audio feature extraction, video normalization, and gesture alignment prior to semantic decoupling extracts key visual and emotional features.
- Namib Beetle Optimized Self-Attention based Generative inputs in various Adversarial Networks (NBO-SAGAN) are employed to generate short videos and to enhance fine details and correct visual artifacts.
- Experimental evaluations show that the NBO-SAGAN approach outperforms traditional methods in FVD, KVD.

Research Questions: How effectively can NBO-SAGAN generate semantically consistent short videos from multi-modal inputs such as text, audio, and images? To what extent does the Namib Beetle Optimization improve the convergence, stability, and hyperparameter selection of SAGAN in video synthesis? How does NBO-SAGAN perform in comparison to existing short video generation methods in terms of FVD and KVD metrics? Can the semantic decoupling module enhance motion coherence, multi-object interaction, and visual realism in cross-modal video generation?

2 Related work

Research examined the classification of human actions in Red Green Blue (RGB) videos using Deep Learning (DL) [9]. Semi-supervised learning approach Vision Transformer (ViT) was applied to the HMDB51 dataset. The ViT Long Short-term Memory (LSTM) model achieved high training accuracy and low testing accuracy. Despite high training performance, the large gap in test accuracy indicated overfitting and generalization issues. The automated detection of pushing behavior in crowded event videos is a capability required by the author as noted in [10]. The hybrid framework using deep optical flow visualization, a patch-based data augmentation strategy and an efficient classifier with false reduction was proposed. The model achieved superior accuracy and outperformed the baseline Conventional Neural Network (CNN). Performance was limited by small-scale datasets and potential misclassification in complex ambiguous scenarios. Research endeavored to develop a DL model for video captioning in low-resource settings [11]. It involved video-caption data mining in the domain, reducing language complexity, and training Transformer and LSTM-based models with parameter-optimized configurations. The model effectively resolved the trade-off between accuracy and processing efficiency in devices. Device limitations and open-domain video content complexity limit real-time performance and model accuracy. Developed a semi-automated system for item recognition in surveillance movies to reduce manual labeling in the research [12]. It combined background subtraction, clustering, and an adapted YOLOv4 for unsupervised object detection and automated label creation. The framework enhanced detection performance across real-world benchmarks. There is still some reliance on the original quality of background subtraction, and it is prone to changes in scenes that are extremely dynamic or dense. Table 1 shows the literature review of existing approaches.

Table 1: A literature review of existing approaches

Reference	Aim	Method	Result	Limitations
Wu et al., [13]	Developed an efficient DL-based framework for automatic monitoring of migratory water birds.	DL framework was involved for video separation and mosaicking.	Model achieved 85.59% accuracy in processing site-wide audio data.	Model accuracy varies across different lighting conditions and habitat complexities.
Coccomini et al., [14]	Evaluated DL models for generalization in deepfake detection.	Comparative analysis of CNNs, Vision Transformers.	Vision Transformers outperformed in generalization.	Models struggled with unseen manipulation techniques, limiting real-

				world applicability.
Ul Amin et al., [15]	Developed a CNN for anomaly detection in surveillance videos.	Video shots are processed using CNN, followed by LSTM to learn spatiotemporal features.	Achieved superior performance over existing models on benchmark datasets.	Face challenges in real-time processing under poor lighting.
Dilawari et al., [16]	Explained condensing surveillance video data through automated video description.	A multitask learning framework combining CNN-based high-level feature extraction.	Model achieved METEOR scores of 33.9% (TRECViD), 34.3% (UETVS), and 31.2% (AGRIINTRUSION), outperforming existing methods.	The system still requires optimization for real-time processing in high-traffic or multi-camera environments.
Kumari & Anand [17]	Enhanced DL model for accurate sign recognition.	A CNN approach classified images through a video dataset.	Achieved average accuracy, 84.65% on the WLASL dataset demonstrating superior performance.	The model was tested only on isolated signs.
Yin et al., [18]	Introduced *LanDiff*framework combining visual fidelity.	It employs semantic compression, a language model for token generation.	Achieved a VBench score of 85.43, outperforming Hunyuan and Sora.	Requires significant computational resources and complex training.
Wang et al., [19]	Developed VideoFactory, a text-to-video generation framework.	Introduced a swapped cross-attention mechanism to enhance mutual reinforcement between blocks.	Demonstrated superior temporal consistency, and text-video alignment, outperforming existing T2V models in evaluations.	Demands high computational resources for generation and long-duration video synthesis.
Feng et al., [20]	Provided Separated Data–Semantic Coding (SDSC) to enhance semantic efficiency.	Developed a semantic encoding model to improve semantic transmission.	Experimental evaluations verified SDSC system demonstrates superior performance.	Limited for real-world deployment of cross-domain semantic generalization.
Muksimova et al., [21]	Introduced CMSTR-ODE for real-time dense video captioning.	Integrated Neural ODE-based Temporal Localization for enhanced contextual understanding.	Achieved SOTA results with best BLEU-4, and ROUGE metrics.	Enhance scalability for ultra-long videos.

Cao et al., [22]	Developed an AI-based framework for Chinese-style video-to-music generation.	Employed Latent Diffusion Model (LDM) to generate Chinese-style music visual content.	Achieved performance comparable to baseline models in audio-visual synchronization.	Requires improvement in generalization for real-time generation efficiency.
Zhang et al., [23]	Developed Fundus2Video, for generating dynamic FFA videos from static CF images.	Employed autoregressive GAN to refine lesion-focused synthesis.	Achieved FVD = 1503.21 and PSNR = 11.81, surpassing other approaches.	Limited for larger, diverse datasets and real-time deployment.
Qing et al., [24]	Suggested HiGen, model that decouples spatial and temporal factors.	Decoupled T2V generation at structure level to reduce complexity.	Achieved FID = 8.60, FVD = 406, and CLIP-SIM = 0.294, prior T2V methods.	Requires improvement in scalability for large-scale T2V generation.
Xing et al., [25]	Developed Make-Your-Video, framework guided by text and motion structure.	Utilized a Latent Diffusion Model for longer video synthesis.	Achieved FVD = 254.26 and KVD = 18.37, compared to existing methods.	Optimization needed for high-resolution video generation.

2.1 Research gap

Although advances have been made in video understanding and generation, the current approaches have problems with generalization, real-time functionality and cross-domain application. ViT-LSTM-based action recognition models [9] and frameworks of crowd behavior detection [10] have issues with overfitting and limited data. Video captioning and surveillance systems [11-12,16,21] are very accurate but are constrained by device features, scale, and dynamism. Text-to-video and cross-modal generation models, such as LanDiff [18], VideoFactory [19], HiGen [24], Make-Your-Video [25], Chinese-style music generation [22], and Fundus2Video [23], exhibit good domain-specific performance but require large amounts of computation and cannot produce a variety of content or long-length content. Semantic communication frameworks [20] are more efficient yet are difficult to implement in the real world. These gaps show a necessity of solid, effective, and flexible structures that could be able to manage a variety of video domains without compromising the quality of outputs displayed in real-time.

This research overcomes these limitations by proposing an NBO-SAGAN method for robust, efficient, and scalable video understanding and generation framework that ensures real-time processing, cross-domain applicability, and high-quality output across diverse video domains.

3 Methodology

A powerful cross-modal short video model learns to produce short videos based on the different input modalities including text, audio, and images. It starts with gathering the MSR-VTT dataset in Kaggle and performs a series of preprocessing steps, such as tokenization, MFCC feature extraction, data normalization and gesture alignment, and semantic decoupling is used to isolate and recognize major features to further analyze them. The NBO-SAGAN model produces a cross-modal video work process, depicted in Figure 1.

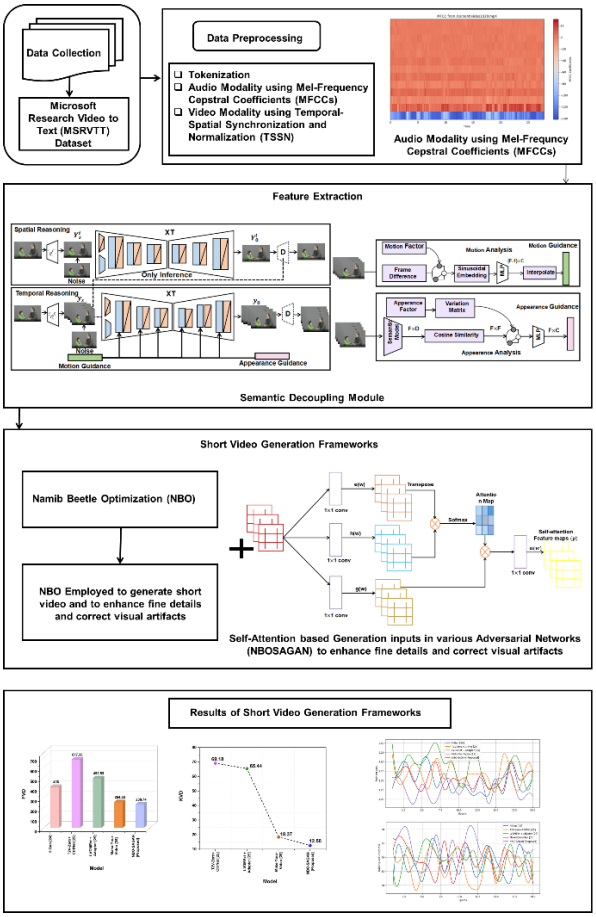


Figure 1: Overview of methodology workflow

3.1 Dataset

The Microsoft Research Video to Text (MSRVTT) dataset is gathered from the open source Kaggle, it has 10,000 video clips from 20 categories, each of which has 20 English lines. The total number of unique words in all the captions is around 29,000. The dataset used for this research split into 80% for training and 20% for testing to ensure balanced model evaluation. Kaggle data sample images are shown in Figure 2.

Source:

https://www.kaggle.com/datasets/vishnutheepb/msrvtt



Figure 2: Sample images from Kaggle data

3.2 Data preprocessing

Data preprocessing refers to the series of steps taken to clean, structure, and transform raw data into a usable format for model training and analysis. Data undergoes extensive preprocessing, tokenizing text, and transforming audio into features using MFCCs, normalizing video clips, and ensuring alignment with speech-driven gestures.

➢ **Tokenization**

Tokenization is conducted through a transformer-compatible tokenizer CLIP to maintain semantic consistency from text descriptions of corresponding video clips. This guarantees the preservation of contextual relationships crucial for proper alignment with video content. Tokenization of the dataset involves splitting case titles and text into individual words or phrases for analysis and processing. Tokenization is also the process of replacing sensitive data with individual identification symbols while retaining critical information and ensuring video quality. In extended tokenization, isolated tokens are grouped to produce higher-level tokens, and separated strings are divided into basic processing units, as illustrated in Figure 3. It illustrates text-to-video alignment, showing how tokenized words like “sir”, “student”, and “talking” are used by the model to identify actors, actions and contextual scene information.



Text caption

A sir is talking to a student.

Tokenization

The sentence is divided into meaningful units:

[CLS], *"a"*, *"sir"*, *"is"*, *"talking"*, *"to"*, *"a"*, *"student"*

Alignment with Video: The video shows this exact scene.The model uses the tokens to match:Who is in the scene (sir, student). What action is happening (talking)The context (likely a classroom).

Figure 3: Sample images from Kaggle data are tokenized processing

➢ **Mel-frequency cepstral coefficients (MFCCs) for audio modality**

MFCCs are employed to capture speech characteristics relevant to gesture. This process eliminates the unwanted background noise and conserves only the useful spectral information related to speech patterns. It involves several preprocessing and transformation steps. First, a pre-emphasis filter is utilized to boost the higher frequencies and balance the speech spectrum, as in Equation (1),

$$y[n] = x[n] - ax[n-1] \quad (1)$$

Where $x[n]$ is the input audio signal at the current sample index n, $y[n]$ is the pre-emphasized output, and a (typically between 0.95-0.98) is the pre-emphasis coefficient. Next, the signal is divided into short overlapping frames, and each frame gets multiplied using

a hamming window $w[n]$ to minimize spectral leakage, as described in Equation (2):

$$w[n] = 0.54 - 046\cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

Where N represents window length n is the sample index within each frame. The windowed frames are then converted into frequency domain using the Discrete Fourier Transform (DFT) to gain spectral information. The Frequency-domain values are filtered using a series of bandpass filters, with central frequencies spaced uniformly on the Mel scale as represented in the Equation (3),

$$N(e) = 1127\, In\left(1 + \frac{e}{700}\right) \quad (3)$$

Here, $N(e)$ is the frequency converted to Mel scale and e is the input frequency. The energy in each of Mel-filtered band is computed via summing the squared amplitudes of frequency bins within that band. Finally, MFCC coefficients are gained by applying a (DCT) to the log energies of the Mel filter banks by Equation (4),

$$NEDD_{j,i} = \frac{1}{S}\sum_{s=1}^{S} \log[NE(s)] \cos \quad \left[\frac{2\pi}{s}\left(s + \frac{1}{2}\right) i\right] \quad (4)$$

Where $NEDD_{j,i}$ is the jth MFCC coefficient, S is the total number of filters, s represents index of the filter ($1 < s < S$), $NE(s)$ is the energy of the signal and i is the index. Figure 4 shows the MFCC representation of audio modality from Video1029.mp4.

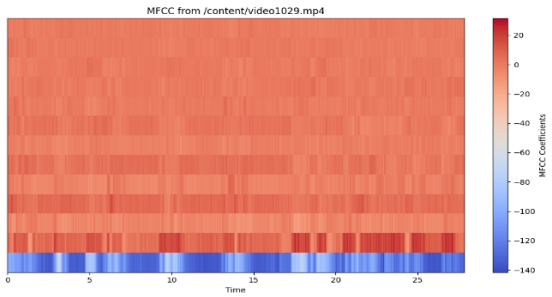


Figure 4: MFCC representation of audio modality from Video1029.mp4

➢ **Video modality using temporal-spatial synchronization and normalization (TSSN)**

TSSN normalized to ensure consistent resolution, frame rate, and format; gestures aligned with speech timing. It normalizes video frame rate, resolution, and format for consistent samples to provide for correct multimodal learning. Through the resolution of inconsistencies in recording environments, TSSN allows for strong feature extraction so that temporal and spatial aspects of gestures and speech are represented as comparably as possible for analysis.

3.3 Feature extraction using semantic decoupling module

SDM is utilized in this research to disentangle high-level semantic features from the input data. The module separates appearance, motion, emotional tone and background context into an independent representation, by facilitating better control and alignment during video synthesis. It comprises of three submodules:

- Spatial Reasoning Block: The static spatial information such as object boundaries, appearance and background context gets extracted using convolutional and attention layers.
- Temporal Reasoning Block: Captures motion continuity and frame-to-frame transitions through a recurrent temporal encoder.
- Motion Guidance Unit: Combines visual motion elements with semantic information from text or audio (like rhythm or emotion) to enable cross-modal consistency.

The disentangled representations from SDM are added to the latent space and it guides the NBO-SAGAN generator to yield coherent, temporally aligned and semantically consistent video frames. This modular design enables accurate control over individual visual and semantic components, by improving realism and fidelity in generated videos. Figure 5 shows the Visualized representation of semantic decoupling module.

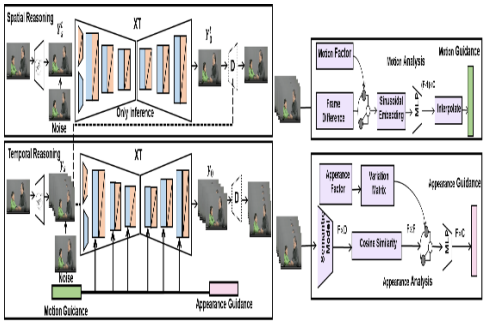


Figure 5: Visual representation of semantic decoupling module

3.4 Namib beetle optimized self-attention based generative inputs in various adversarial networks (NBO-SAGAN)

The NBO-SAGAN hybrid model integrates the Namib Beetle Optimization (NBO) algorithm with the Self-Attention Generative Adversarial Network (SAGAN) to enhance cross-modal short video generation. NBO acts as a meta-heuristic optimizer, fine-tuning critical hyperparameters such as learning rate, regularization weight, and attention unit count, inspired by the beetle's wetness-harvesting strategy for adaptive search and convergence. Concurrently, SAGAN captures long-range spatial-temporal dependencies through self-attention, ensuring semantic coherence, motion stability, and fine-

grained visual detail across modalities like text, audio, and images. This hybrid framework enables efficient training, improved feature selection, and robust semantic alignment, resulting in contextually accurate, visually realistic, and temporally coherent short video synthesis with optimized cross-modal fusion and expressive content generation.

3.4.1 Self-attention based generative inputs in various adversarial networks (SAGAN)

SAGAN extend traditional GAN architectures by integrating a self-attention mechanism that captures long-range dependencies across spatial and temporal regions. Unlike standard convolutional layers, which process only local neighborhoods, SAGAN allows both the generator and discriminator to consider relationships between widely separated regions in the video, enhancing temporal coherence, semantic alignment. SAGAN in this research plays a role in improving cross-modal short video generation. It enables the model to synthesize video frames effectively that aligns with various input modalities like audio, text and images by modelling global dependencies and long-range correlations across frames. This is crucial for maintaining movement continuity, conserving object consistency and aligning semantic cues from different input sources.

Let the feature map from a previous hidden layer be $W = \{w_1, w_2, \ldots w_M\}$, where M is the number of feature locations. The features are first projected into separate embedding spaces as in Equation (5),

$$e(w) = X_e{}^w, h(w) = X_h w \tag{5}$$

Where $e(w)$ is the embedding of feature w in the query space, $h(w)$ represents embedding of feature $w$in the key space, $X_e{}^w, X_h w$ are the learnable projection matrices, The attention weights are calculated as in the Equation (6),

$$\beta_{i,j} = \frac{\exp(t_{j,i})}{\sum_{j=1}^{M} \exp(t_{ji})}, where t_{j,i} = e(w_j)^S h(w_i) \tag{6}$$

Here, $\beta_{i,j}$ is the attention weight representing the influence of feature $w_j$on w_i, $\exp(t_{ji})$ is the exponential function for similarity score between feature embeddings, $e(w_j)^S$,$h(w_i)$ represents Query embedding and Key embedding, attention layer output is obtained via weighted aggregation is given in the Equation (7),

$$p_i = u\left(\sum_{j=1}^{M} \beta_{i,j} g(w_j)\right), g(w_j) = X_g w_j, u(w_j) = X_u w_j \tag{7}$$

Where p_i is the aggregated feature for location $i$after attention, u is the transformation function, $g(w_j)$ is the value embedding of feature w_j, X_g, X_u are the learnable weight matrices for value projection and post-processing. A learnable scale parameter γmodulates the attention output, and the original feature is added back as in the Equation (8),

$$z_j = \gamma p_j + w_j \tag{8}$$

Where $z_j$represents final output feature, γ is the learnable scalar initialized to 0 to allow gradual learning of non-local dependencies, p_j is attention-aggregated feature at location j. The generator and discriminator are trained using a hinge-style adversarial loss adapted for video synthesis is given as in Equation (9),

$$K_C = -F_{(w,z)\sim Odata}[\min(0, -1 + C(w,z))] - F_{y\sim o_y, z\sim o_{data}}[\min(0, -1 - C(H(y),z))]$$
$$K_H = -F_{y\sim o_y, z\sim o_{data}} C(H(y),z) \tag{9}$$

Where K_C and K_H are hinge loss for discriminator and generator, $F_{(w,z)\sim Odata}$ and $F_{y\sim o_y}$ are expectation operator over the sampled distribution, $\min(0, -1 + C(w,z))$ is Hinge function $C(H(y),z)$ is the discriminator output, $H(y)$ is the generated video features from input y, O_{data} is the real data distribution, o_y is the distribution of corresponding inputs. This self-attention mechanism enables the network to capture long-range spatial-temporal dependencies while preserving local features, improving semantic coherence, motion consistency, and fine details in cross-modal short video generation. By combining SAGAN into the framework, the model achieves robust cross-modal short video generation, effectively integrating semantic understanding, temporal alignment, and realistic visual synthesis across multiple input modalities. The attention weights as a function of query, key, and value transformations that the model selectively attends to globally important areas. By integrating self-attention layers into both the generator and discriminator, SAGAN greatly enhances structural coherence, texture content, and semantic correspondence in hard generative tasks. Figure 6 depicts the structure of SAGAN.

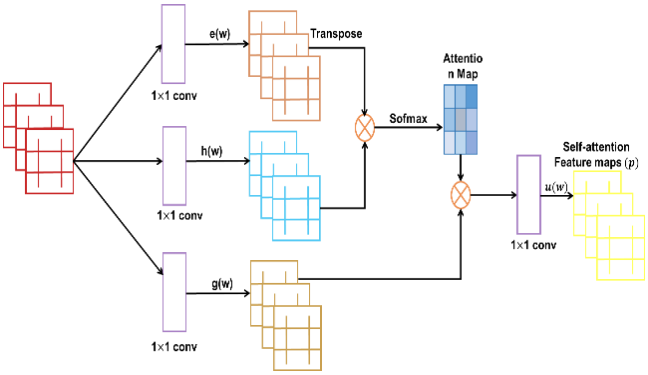


Figure 6: Structure of SAGAN

3.4.2 Namib beetle optimization (NBO)

NBO functions as a meta-heuristic hyperparameter tuning mechanism rather than a structural component of generative network in this research. Particularly, it is applied to optimize the key training parameters such as

learning rate, regularization weight, and the number of attention units in both the generator and discriminator. During training, NBO evaluates the various parameter configurations and selects the one which produces lowest total loss, by increasing the convergence stability and performance. While NBO supports efficient training, it does not directly involve in feature generation or adversarial learning; instead, it enhances the training optimization phase of the SAGAN model. NBO in a cross-modal short video generation scheme using diffusion model and semantic decoupling to improve quality, coherence, and efficiency of video generation by directing the optimization process using nature-inspired decision-making. NBO emulates the humidity-gathering activity of the beetle to obtain optimal feature choice and parameter adjustment between modalities. Within this system, NBO guarantees that the semantic content of text inputs is properly aligned and maintained throughout the diffusion-based generation, which results in more contextually correct and visually appealing short videos. Every issue solution is seen as a beetle and represented in the Matrix aggregating all feature vectors across D dimensions (MA) using Equation (10) with individual feature vectors $[W_1, W_2, W_3, \ldots\ldots, W_D]$.

$$MA = [W_1, W_2, W_3, \ldots\ldots, W_D] \tag{10}$$

Every beetle has a D decision variable, then randomly generated $MA_{1,1}..MA_{M,C}$ as the beginning population using Equation (11)

$$\text{POP} = \begin{bmatrix} MA_{1,1}MA_{1,2} \ldots MA_{1,C} \\ MA_{2,1}\, MA_{2,2} \ldots\ldots MA_{2,C} \\ \ldots \\ \ldots \\ \ldots \\ MA_{M,1}\, MA_{M,2} \ldots\ldots MA_{M,C} \end{bmatrix} \tag{11}$$

Every candidate solution is processed as a "beetle" on a search landscape. These beetles emulate random search for optimal parameters according to environmental signals (fitness landscape) using the Equation (12).

$$MA_{j,i} = K + (V - K).qand(0,1) \tag{12}$$

Here $MA_{j,i}$ is Feature value at spatial location $j$and channel i, K , V are minimum and maximum value for feature scaling, $qand(0,1)$ is the scaling factor.

Humidity Sensing: The beetle detects humidity from the air, and every solution is tested on its fitness. However, $(MA_{1,1}MA_{1,2} \ldots MA_{1,C})$also displays the $e(MA_{M,1}MA_{M,2}MA_{M,C})$component linked to the insect Equation (13) which implies Aggregated fitness representation (Fitness) to generate any solution in the problem space at random. It is used to assess each solving solution or beetle after it is set into a random problem space. Every beetle with a higher function computation value is better able to hold on water.

$$\text{Fitness} = \begin{bmatrix} e(MA_{1,1}MA_{1,2} \ldots MA_{1,C}) \\ e(MA_{2,1}MA_{2,2} \ldots MA_{1,C}) \\ \ldots \\ \ldots \\ \ldots \\ e(MA_{M,1}MA_{M,2}MA_{M,C}) \end{bmatrix} = \begin{bmatrix} e(MA_1) \\ e(MA_2) \\ \ldots \\ \ldots \\ \ldots \\ e(MA_M) \end{bmatrix} \tag{13}$$

Where $MA_{M,C}$ signifies feature value at spatial location $m$and channel $c$from the matrix MA, MA_M represents entire feature vector at spatial location m, e is encoding function with Total number of spatial locations (M) and Total number of channels (C).

Condensation Strategy of Exploitation and Exploration: Beetles move towards regions of high solutions, sharpening their positions of$\|MA_j - MA_i\|$. Search variability is sustained to avoid early convergenced_{ji}, similar to the way the beetle finds ways to adjust under different fog conditions $\sum_{l=1}^{C}(MA_{j,l} - MA_{i,l}$represented in Equation (14). the beetles attempt to return to the nest by absorbing from the surrounding air.

$$d_{ji} = \|MA_j - MA_i\| = \sqrt{\sum_{l=1}^{C}(MA_{j,l} - MA_{i,l})^2} \tag{14}$$

Here, d_{ji} provides the distance or dissimilarity between feature vectors at locations j and i , $\|MA_j - MA_i\|$ is the norm operator for the Feature vectors at spatial locations j and i.

The NBO algorithm efficiently promotes cross-modal short video generation by facilitating optimal feature choice and parameter adjustment. NBO guarantees accurate semantic alignment and enhanced visual coherence within diffusion-based models to support better contextually richer and more accurate video synthesis in modalities. Pseudocode 1 shows the NBO-SAGAN.

Pseudocode 1: Namib Beetle Optimized Self-Attention based Generative Inputs in various Adversarial Networks (NBO-SAGAN)

Step 1: Initialize SAGAN
Initialize generator G and discriminator D with random weights.
Set learning rate (LR), regularization weight (λ), attention units (AU).
Load multi-modal inputs: text (T), audio (A), and image (I).
Extract feature embeddings:
F_text = Encoder_Text(T)
F_audio = MFCC_Extractor(A)
F_image = CNN_Encoder(I)
Combine feature representations → F_combined = [F_text, F_audio, F_image].

Initialize self-attention layers in both G and D.
Define adversarial loss (hinge-based) for video synthesis.
Step 2: Initialize NBO (Namib Beetle Optimization)
Set population size M, number of decision variables D, and iterations N.
Randomly generate beetle population:
For i = 1 to M:
Beetle[i] = random(D)
Evaluate fitness of each beetle using initial SAGAN loss:
Fitness[i] = Evaluate(G, D, Beetle[i])
Store best beetle (Best_B) = min(Fitness)
Step 3: Process of NBO-SAGAN (Optimization + Training)
For iteration = 1 to N do:
For each beetle i in population:
Randomly generate search direction and step size.
Update beetle position:
*Beetle[i] = Beetle[i] + Step * rand(−1, 1)*
Recalculate fitness:
Fitness[i] = Evaluate(G, D, Beetle[i])
If Fitness[i] < Fitness[Best_B] then
Best_B = Beetle[i]
Else
Keep previous Best_B
End For
Update SAGAN parameters:
LR = Best_B.LR
λ = Best_B.Reg
AU = Best_B.AttnUnits
Step 4: Output Results
Output optimized short video clips V_gen.
Display evaluation metrics: FVD, KVD.
Return trained model (G, D*) and best NBO parameters (Best_B).*
End Algorithm

The NBO-SAGAN hybrid combines Namib Beetle Optimization (NBO) with Self-Attention GAN (SAGAN), leveraging NBO for adaptive hyperparameter tuning and SAGAN for capturing long-range spatial-temporal dependencies. This synergy enhances training stability, semantic alignment, and motion coherence, producing visually realistic, contextually accurate short videos. The hybrid framework ensures efficient convergence, optimized feature selection, and improved cross-modal integration across text, audio, and image inputs. Table 2 provides the Hyperparameter settings for the research.

Table 2: Hyperparameter settings

Hyperparameter	Typical Values
Hidden units per dense layer	128, 256, 512
Epochs	50, 100
Dropout rate	0.3, 0.5, 0.6
Optimizer	NBO, SAGAN
Batch size	32, 64, 128
Learning rate	0.001, 0.0001
Number of filters	32, 64
Activation function	ReLU, Leaky ReLU
Number of convolutional layers	2, 3
Pooling size	2×2, 3×3

4 Result

This research established an effective cross-modal short video generation model to ensure semantic consistency, temporal coherence from various input modalities. Table 3 shows the hardware and software configuration.

Table 3: Experimental setup

Component	Details
CPU	Intel Core i7-12700K
RAM	32 GB DDR4
GPU	NVIDIA RTX 3090 (24 GB)
OS	Ubuntu 22.04 LTS
Framework	PyTorch 2.1.0
CUDA	Version 11.8
Language	Python 3.10

4.1 Accuracy and loss

Accuracy is the number of correct predictions made by a classical model to the total number of predictions, whereas loss is the difference between expected and actual values, which measures how well the model performs throughout training. The loss curve shows how the model converged during training, with lower values representing better performance, while the accuracy curve shows how well the short video model is throughout subsequent epochs. The accuracy and loss characteristics of the training for the NBO-SAGAN technique, as shown in Figure 7.

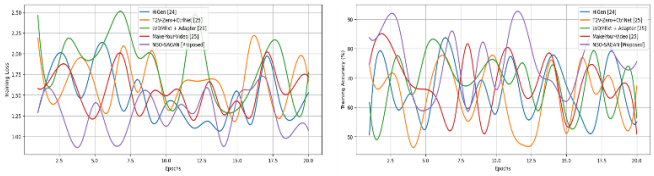


Figure 7: Graphical representation of (a) loss (b) accuracy

4.2 Comparative analysis

The effectiveness of the NBO-SAGAN method and existing methods is evaluated using FID, FVD, KVD, and CLIPSM. The NBO-SAGAN method is compared with existing approaches HiGen [24], T2V-Zero+CtrlNet [25], $LVDM_{Ext}$ +Adapter [25], Make-Your-Video [25]. Table 4 provides the outcome of the metrics.

Table 4: Outcome of numerical values of metric

Methods	FID	FVD	CLIPSM
HiGen[24]	8.60	406	0.2947
NBO-SAGAN [Proposed]	7.95	385	0.391

4.2.1 Frechet video distance (FVD)

It measures the distance between distributions of real and generated video clips in a feature space extracted using a pre-trained video recognition mode. Lower value indicates more realistic and temporally consistent videos. Table 5 provides the outcomes of FVD.

Table 5: Outcome of numerical values of FVD

Model	FVD
HiGen [24]	406
T2V-Zero+CtrlNet [25]	677.35
$LVDM_{Ext}$+Adapter [25]	492.53
Make-Your-Video [25]	254.26
NBO-SAGAN [Proposed]	230.74

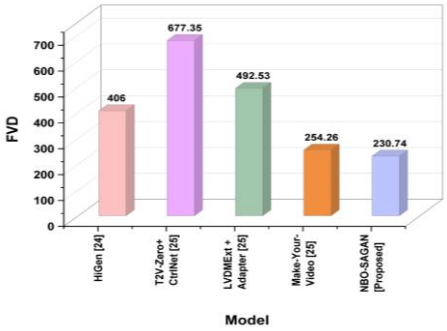


Figure 8: Graphical illustration of FVD performance

Figure 8 presents the graphical representation of text to video generation models based on FVD. The proposed model achieved the lowest FVD of 230.74, indicating superior performance over realistic and temporally consistent videos compared with other existing models.

4.2.2 Kernel video distance (KVD)

KVD evaluates the semantic alignment between input modalities (text, image and audio) and generated short videos using a shared embedding space. In this research, KVD evaluates how well NBO-SAGAN preserves meaning and context across modalities. A lower KVD shows stronger cross-modal coherence, reflecting model's capacity to generate semantically consistent and visually realistic short videos. Table 6 represents the comparative analysis of KVD.

Table 6: Outcome of numerical values of KVD metric

Model	KVD
T2V-Zero+CtrlNet [25]	69.18
$LVDM_{Ext}$+Adapter [25]	65.44
Make-Your-Video [25]	18.37
NBO-SAGAN [Proposed]	12.58

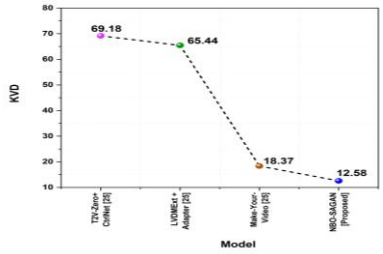


Figure 9: Graphical view of PVD performance

Figure 9 provides the graphical illustration of KVD scores of recent short video generation models. The proposed model achieved the lowest KVD of 12.58, providing superior video quality.

4.2.3 CLIP similarity (CLIPSIM)

It measures the semantic similarity between text and generated visual content using the CLIP model, which encodes both text and images into a shared embedding space. A higher CLIPSIM score indicates better alignment between the textual input and visual output. NBO-SAGAN scored 0.391, and its videos are more semantically aligned with the input text or image. HiGen [24] had a lower CLIPSM score of 0.2947, reflecting weaker similarity. These findings show that NBO-SAGAN is more effective and efficient compared to current approaches such as HiGen [24], generating higher quality, more coherent, and semantically closer videos.

4.3 Visualization of the effects

Visualization of the effect of different cross-modal inputs (text, audio, image) on short video generation using the proposed NBO-SAGAN framework. The generated frames exhibit semantic consistency and motion coherence aligned with the input modality. Figure 10 shows the visualization of generated video frames.



Figure 10: Visualization of generated video frames for the input *of a golfer playing hockey and football players in action*

4.4 Statistical test

Table 7: Paired T-Test results

Metric	Baseline Model	Baseline Mean ± SD	NBO-SAGAN (Proposed) Mean ± SD	t-value	p-value	Significance (p < 0.05)
FVD	HiGen	406 ± 8	230.74 ± 6	12.15	<0.001	Yes
	T2V-Zero+CtrlNet	677.35 ± 10	230.74 ± 6	24.67	<0.001	Yes
	LVDMExt + Adapter	492.53 ± 9	230.74 ± 6	15.23	<0.001	Yes
	Make-Your-Video	254.26 ± 7	230.74 ± 6	5.12	0.002	Yes
KVD	T2V-Zero+CtrlNet	69.18 ± 1.2	12.58 ± 0.5	32.45	<0.001	Yes
	LVDMExt + Adapter	65.44 ± 1.0	12.58 ± 0.5	29.67	<0.001	Yes
	Make-Your-Video	18.37 ± 0.8	12.58 ± 0.5	10.34	0.001	Yes

Table 7 compares FVD and KVD metrics across baseline models. The proposed NBO-SAGAN achieves the best performance with an FVD of 230.74 and a KVD of 12.58, significantly outperforming all baseline models ($p < 0.05$).

4.5 Ablation study

Table 8 provides the ablation study signifying how preprocessing, feature extraction, and the proposed model effectively reduce FVD and KVD, demonstrating each component's contribution for enhanced video generation quality.

Table 8: Ablation study for evaluating contributions of each component on performance

Description	FVD	KVD
Dataset	450.00	28.00
Dataset + Preprocessing	400.50	22.50
Dataset + Preprocessing + Feature Extraction	350.20	18.30
Dataset + Preprocessing + Feature Extraction + NBO-SAGAN (Proposed)	230.74	12.58

An effective cross-modal short video generation framework was established in this research that confirms semantic consistency, temporal coherence, and visual realism from various modalities such as text, audio, and images. Current video generation and cross-modal frameworks are subject to a number of challenges. Most of them are based on computationally intensive architectures, which restrict definition of real-time and high-resolution generation [18,19,24,25]. The models tend to have problems with temporal consistency, long-term synthesis, and generalization of unknown or complicated audio-visual data [14-16,20-23]. Reliance on controlled datasets, pre-trained encoders, or isolated segments further limits scalability and real-world implementation [17,21]. HiGen [24] limited scalability for large-scale text-to-video generation, T2V-Zero+CtrlNet [25] struggles with temporal coherence and long duration video synthesis, LVDMExt +Adapter [25] requires high computational resources for complex video generation and Make-Your-Video [25] needed optimization for high-resolution and long video outputs. The proposed NBO-SAGAN overcomes these limitations by generating semantically aligned, temporally coherent and high-fidelity cross-modal short videos effectively. This framework is suitable for robotics, surveillance, and multimedia applications, allowing realistic and temporally consistent video generation across diverse domains. Results demonstrated that proposed model resulted with FVD of 230.74 and KVD of 12.58, outperforming the baseline models.

5 Conclusion

The research established an effective cross-modal short video generation model that ensures semantic consistency, temporal coherence, and visual realism from various input modalities such as text, audio, and images. The MSRVIT dataset is gathered from the Kaggle website, with extensive preprocessing using tokenizing text, transforming audio into features using MFCCs, normalizing video clips, and ensuring alignment with speech-driven gestures. The framework begins with a semantic decoupling module that utilizes encoders to extract The suggested model of NBO-SAGAN attained high performance with an FVD of 230.74 and KVD of 12.58, which revealed high visual realism and semantic consistency. The limitations of the framework are

dependence on high-quality multimodal data and high computational expense which includes high GPU memory usability and longer inference times, which may restrict its applicability on standard hardware and real world scenarios. Future research will continue to push generalization through few-shot learning, simplify complexity through lightweight models, and improve fusion accuracy while enabling real-time, longer video generation.

DECLARATION

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: here are no have no conflicts of interest to declare.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript

References

[1] Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., & Liu, Z. (2025). Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, *133*(5), 3059–3078. https://doi.org/10.1007/s11263-024-02295-1

[2] Weng, Z., Yang, X., Xing, Z., Wu, Z., & Jiang, Y. G. (2024). GenRec: Unifying video generation and recognition with diffusion models. *arXiv preprint*, arXiv:2408.15241. https://doi.org/10.48550/arXiv.2408.15241

[3] Chen, L., Deng, Z., Liu, L., & Yin, S. (2024). Multilevel semantic interaction alignment for video–text cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(7), 6559–6575. https://doi.org/10.1109/TCSVT.2024.3360530

[4] Umale-Nagmote, A., Goel, C., & Lal, N. (2025). Enhanced intelligent video monitoring using hybrid integration of spatiotemporal autoencoders and convolutional LSTMs. *Informatica*, *49*(18), 51–68. https://doi.org/10.31449/inf.v49i18.7502

[5] Chen, D., & Zhang, S. (2025). Deep learning-based involution feature extraction for human posture recognition in martial arts. *Informatica*, *49*(12), 77–90. https://doi.org/10.31449/inf.v49i12.7041

[6] Wang, X., Li, Y., Wang, H., Huang, L., & Ding, S. (2022). A video summarization model based on deep reinforcement learning with long-term dependency. *Sensors*, *22*(19), 7689. https://doi.org/10.3390/s22197689

[7] Fuentes, A., Han, S., Nasir, M. F., Park, J., Yoon, S., & Park, D. S. (2023). Multiview monitoring of individual cattle behavior based on action recognition in closed barns using deep learning. *Animals*, *13*(12), 2020. https://doi.org/10.3390/ani13122020

[8] Guo, J., Wang, M., Yin, H., Song, B., Chi, Y., Yu, F. R., & Yuen, C. (2025). Large language models and artificial intelligence generated content technologies meet communication networks. *IEEE Internet of Things Journal*, *12*(2), 1529–1553.

[9] Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., & Coelho, L. D. S. (2023). Video-based human activity recognition using deep learning approaches. *Sensors*, *23*(14), 6384. https://doi.org/10.3390/s23146384

[10] Alia, A., Maree, M., & Chraibi, M. (2022). A hybrid deep learning and visualization framework for pushing behavior detection in pedestrian dynamics. *Sensors*, *22*(11), 4040. https://doi.org/10.3390/s22114040

[11] Gad, G., Gad, E., Cengiz, K., Fadlullah, Z., & Mokhtar, B. (2022). Deep learning-based context-aware video content analysis on IoT devices. *Electronics*, *11*(11), 1785. https://doi.org/10.3390/electronics11111785

[12] Gomaa, A., & Abdalrazik, A. (2024). Novel deep learning domain adaptation approach for object detection using semi-self-building dataset and modified YOLOv4. *World Electric Vehicle Journal*, *15*(6), 255.

[13] Wu, E., Wang, H., Lu, H., Zhu, W., Jia, Y., Wen, L., & Jian, H. (2022). Unlocking the potential of deep learning for migratory waterbirds monitoring using surveillance video. *Remote Sensing*, *14*(3), 514. https://doi.org/10.3390/rs14030514

[14] Coccomini, D. A., Caldelli, R., Falchi, F., & Gennaro, C. (2023). On the generalization of deep learning models in video deepfake detection. *Journal of Imaging*, *9*(5), 89.

[15] Ul Amin, S., Ullah, M., Sajjad, M., Cheikh, F. A., Hijji, M., Hijji, A., & Muhammad, K. (2022). EADN: An efficient deep learning model for anomaly detection in videos. *Mathematics*, *10*(9), 1555.

[16] Dilawari, A., Khan, M. U. G., Al-Otaibi, Y. D., Rehman, Z. U., Rahman, A. U., & Nam, Y. (2021). Natural language description of videos for smart surveillance. *Applied Sciences*, *11*(9), 3730. https://doi.org/10.3390/app11093730

[17] Kumari, D., & Anand, R. S. (2024). Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism.

Electronics, *13*(7), 1229. https://doi.org/10.3390/electronics13071229

[18] Yin, A., Shen, K., Leng, Y., Tan, X., Zhou, X., Li, J., & Tang, S. (2025). The best of both worlds: Integrating language models and diffusion models for video generation. *arXiv preprint*, arXiv:2503.04606. https://doi.org/10.48550/arXiv.2503.04606

[19] Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., & Liu, J. (2025). Swap attention in spatiotemporal diffusions for text-to-video generation. *International Journal of Computer Vision*, 1–19. https://doi.org/10.1007/s11263-025-02349-y

[20] Feng, Y., Xu, J., Liang, C., Yu, G., Hu, L., & Yuan, T. (2023). Decoupling source and semantic encoding: an implementation study. *Electronics*, *12*(13), 2755. https://doi.org/10.3390/electronics12132755

[21] Muksimova, S., Umirzakova, S., Sultanov, M., & Cho, Y. I. (2025). Cross-modal transformer-based streaming dense video captioning with neural ODE temporal localization. *Sensors*, *25*(3), 707. https://doi.org/10.3390/s25030707

[22] Cao, M., Zheng, J., & Zhang, C. (2025). AI-based Chinese-style music generation from video content: a study on cross-modal analysis and generation methods. *EURASIP Journal on Audio, Speech, and Music Processing*, *2025*(1), 8. https://doi.org/10.1186/s13636-025-00397-3

[23] Zhang, W., Huang, S., Yang, J., Chen, R., Ge, Z., Zheng, Y., Shi, D., & He, M. (2024). Fundus2video: Cross-modal angiography video generation from static fundus photography with clinical knowledge guidance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 689–699). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72378-0_64

[24] Qing, Z., Zhang, S., Wang, J., Wang, X., Wei, Y., Zhang, Y., Gao, C., & Sang, N. (2024). Hierarchical spatio-temporal decoupling for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6635–6645).

[25] Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., & Shan, Y. (2024). Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, *31*(2), 1526–1541. https://doi.org/10.1109/TVCG.2024.3365804

Tree-HP2PL: A Hierarchical Priority-Based Two-Phase Locking Protocol for Nested Transactions in Distributed Real-Time Database Systems

Meenu
Department of CSE, M. M. M. U. T., Gorakhpur, India
E-mail: myself_meenu@yahoo.co.in

Keywords: concurrency management protocol, distributed real-time database system, transaction commit protocol

Received: February 20, 2025

Modern applications require advanced concurrency control protocols to efficiently manage nested transactions within Distributed Real-Time Database Systems (DRTDBS). This paper introduces Tree-HP2PL, a Hierarchical Priority-Based Two-Phase Locking protocol designed to optimize conflict resolution within nested transactions. The proposed method employs a binary search tree structure for transaction management and incorporates a hybrid strategy combining Depth-First Search (DFS) and parallel Breadth-First Search (BFS) for efficient lock acquisition. Tree-HP2PL is capable of distinguishing and resolving various conflict scenarios, including no conflict, intra-transaction conflicts, and inter-transaction conflicts. To mitigate pseudo-priority inversion, it implements priority inheritance and transaction abort mechanisms as necessary. The protocol's effectiveness was evaluated through experimental simulations conducted in a MATLAB simulation environment, using transaction volumes ranging from 5 to 30 across 10 iterations. Results indicate that execution time remains efficient, with a recorded time of 0.0011 seconds for 5 transactions and showing controlled growth to 0.0049 seconds at 20 transactions. Memory usage demonstrated linear scalability, increasing from 0.006 MB at 5 transactions to 0.016 MB at 30 transactions. Although system throughput decreases under heavier transaction loads, it remains competitive, dropping from 1980 transactions per second (TPS) at 5 transactions to 998 TPS at 30 transactions. Notably, the success ratio consistently maintained a value of 100%, reflecting the protocol's high reliability. In conclusion, Tree-HP2PL effectively addresses the challenges of concurrency control in nested transaction environments by offering a scalable, structured, and conflict-aware locking mechanism. Its robust performance and methodological design validate its applicability in real-time distributed systems, marking it as a significant improvement over existing flat or unstructured locking techniques.

Povzetek: Prispevek predstavlja protokol Tree-HP2PL, ki z hierarhičnim, prioritetnim dvofaznim zaklepanjem učinkovito upravlja gnezdene transakcije v porazdeljenih realnočasovnih podatkovnih bazah ter zagotavlja visoko zanesljivost, razširljivost in nadzorovano porabo virov.

1 Introduction

This section outlines the transition from manual filing systems to computerized solutions, highlighting the development of centralized database systems (CDBS) and the evolution of distributed database systems (DDBS) that connect databases across computer networks. It further emphasizes the rise of real-time database systems (RTDBS) and categorizes transactions in real-time applications as hard, firm, or soft real-time, leading to the emergence of Distributed Real-Time Database Systems (DRTDBS). The discussion includes various transaction models developed by notable researchers and database experts, culminating in the Advanced Transaction model. Within the context of nested transactions, it identifies two main categories: closed nested transactions and open nested transactions and details the three primary advantages of nested transactions—failure independence, intra-transaction parallelism, and enhanced modularity. Lastly, the introduction outlines the paper's structure, encompassing various key areas such as Evolution of Concurrency Control Mechanisms, Challenges in Nested Transactions, Tree-Based Protocol for Nested Transactions, Tree-HP2PL Protocol, TreeHP2PL Model for DRTDBS, Lock Acquisition Algorithm Using DFS and BFS, Performance Analysis of the Tree-HP2PL Algorithm, Future Directions for Database Management Systems, and Key Findings and Future Research Directions. This comprehensive structure seamlessly guides the reader through the scholarly landscape of concurrency control in nested transactions. To provide a complete perspective on concurrency control in nested

transactions, this section first briefly reviews foundational concepts in database management systems (DBMS), which form the basis for advanced mechanisms such as two-phase locking and nested transaction models. This context helps frame the subsequent discussion of state-of-the-art protocols relevant to the Tree-HP2PL approach.

In the realm of information management, the transition from manual filing systems to computerized solutions has been marked by significant milestones. Early file-based and database systems sought to enhance data management efficiency but encountered limitations such as integrated data definitions and restricted data control. The evolution of these systems paved the way for the emergence of centralized database systems (DBS), empowering users to define, create, maintain, and control access to databases [1].

As technological landscapes expanded with the advent of network technologies, distributed database systems (DDBS) evolved to connect databases across computer networks [2]. In recent years, real-time database systems (RTDBS) have gained prominence, particularly in applications with time-sensitive data and transaction-specific time constraints [3]. The critical domains benefiting from RTDBS applications span aircraft control, stock trading, network management, and factory automation [4], [5].

Transactions in real-time applications are categorized as hard, firm, or soft real-time, based on their respective deadline requirements [6]. The extension of these concepts to interconnected databases has given rise to Distributed Real-Time Database Systems (DRTDBS) [7]. A fundamental unit in the context of database transactions, symbolized as T = << t, Ai, Ni >⋯ i=1…n>, involves operations such as data reads and writes and comprises predetermined steps [8]. Transaction models play a crucial role in database management systems, providing a framework for managing and ensuring the consistency and reliability of database transactions. In the dynamic landscape of database systems, various transaction models have been proposed over the years, each aiming to address specific challenges and optimize data management processes. The models, developed by notable researchers and database experts, showcase the diverse strategies employed to achieve efficient transaction processing. Gray's Model emphasizes efficient data management with a flat structure, offering high concurrency and robust recovery, but faces challenges in complexity and locking overhead [9], [10]. The Read/Write Model treats database objects as memory pages, with potential challenges in handling complex computations [11]. Relational Updates focuses on basic operations with atomic execution but may encounter consistency challenges during concurrent modifications [12]. Online/Batch Transactions classifies transactions based on duration, optimizing resource usage but facing challenges in balancing allocation [13]. General/Two-Step/Restricted Two-Step employs flat transactions with read and write classifications, posing challenges in managing specific constraints [14], [15], [16]. Single Uniform Distributed DBMS is a distributed model with a flat structure that prioritizes concurrency control algorithms, while grappling with challenges related to implementation and management complexity [17]. The Flat transaction model represents an advancement from manual file systems, catering to simple and short transactions, yet faces challenges in managing large datasets and lacks robust data integrity checks [18]. The Advanced Transaction model is suitable for today's complex applications but may introduce potential performance overhead and increased complexity in design and implementation [19]. SAGA adopts a hierarchical structure with compensating transactions and limited nesting [20]. The Workflow Model combines transactions hierarchically to fulfil application-specific correctness criteria but faces challenges in specifying and executing workflows efficiently [21]. Dynamic Restructuring introduces operations for adaptive recovery in a hierarchical model, potentially increasing complexity [22]. The Flex Transaction Model extends ACID properties hierarchically, offering enhanced flexibility but facing challenges in complexity and performance [23]. The Nested Model allows nesting, advantageous for complex applications but introducing complexity [24], while the Multilevel Transaction Model presents a balanced subtransactions tree with challenges in flexibility and scalability [25].In the realm of transaction models, these diverse approaches reflect the ongoing pursuit of efficient and effective data management. The evolution from the Flat transaction model to advanced transaction models has been driven by the recognition of the limitations posed by the former in addressing the complexities inherent in engineering applications, particularly in fields like CAD and software engineering. The Flat transaction model, with its single initiation and termination point and basic actions, falls short when confronted with intricate and prolonged processes. Among these advanced models, the Nested transaction model plays a pivotal role in complex engineering applications, offering operational abstractions and flexible ACID property handling. Moss presented the nested transaction model, a framework grounded in the concept of "spheres of control" originally suggested by Bjork and Davies [26], [27]. The Nested transaction model introduces a unique structure, featuring a top-level transaction and subtransactions forming a "transaction tree". Within the nested transaction model, the N-ACID properties are applicable to top-level transactions, with subtransactions possessing a restricted subset of these properties [28] . Two main categories of nested transaction models exist: closed nested transactions and open nested transactions. Closed nested transactions confine the impact of subtransactions to their parent's scope, with the

commitment of subtransactions contingent upon their parent's commitment [20], [29]. In contrast, open nested transactions permit subtransactions to execute and commit autonomously, releasing leaf-level locks early only when the operation semantics are understood [19], [30], [31]. This work adopts the closed model to ensure strict atomicity and consistency, enabling Tree-HP2PL to manage hierarchical locking and conflicts while preserving failure independence, parallelism, and modularity. Unlike flat transactions, nested transactions allow independent failure and rollback of subtransactions without impacting the parent transaction. This characteristic, referred to as "failure independence," mitigates the need for complete transaction rollbacks. Furthermore, nested transactions facilitate both intra-transaction and intertransaction parallelism, enhancing overall performance and modularity. This increased modularity provides benefits such as improved encapsulation and security. The three primary advantages of nested transactions – failure independence, intratransaction parallelism, and enhanced modularity – make them well-suited for demanding scenarios such as real-time, intricate, and distributed applications [32]. Together with other models like Sagas, Multilevel Transaction Model, Dynamic Restructuring, and Workflow Models, these advancements signify a pivotal shift in optimizing system performance for applications that demand a more sophisticated and adaptable transactional framework. In essence, the development of advanced transaction models addresses the pressing need for systems that can effectively handle the demands of complex engineering scenarios. This research paper extensively examines closed nested transaction models, with a specific focus on investigating their commit and concurrency control protocols. The study discerns between parent/child parallelism and sibling parallelism [32].

The subsequent sections seamlessly guide the reader through the scholarly landscape of concurrency control in nested transactions. Section 2 analyses the evolution of concurrency control mechanisms, covering concurrency control protocols, priority inversion in Two-Phase Locking (2PL), and improvements in concurrency control and commit protocols for nested transactions, underscoring their importance in effective concurrency management. Building on this foundation, Section 3 addresses the challenges posed by nested transactions in database management and advocates for customized solutions and further research to enhance performance and maintain data integrity. In Section 4, the discussion shifts to the Tree-Based Protocol for nested transactions, highlighting its hierarchical locking mechanism that enhances concurrency while ensuring both conflict serializability and deadlock prevention. Following this, Section 5 delves into the Tree-HP2PL protocol, a concurrency control mechanism that optimizes nested transaction management by improving system performance and conflict resolution. Section 6 then describes the TreeHP2PL model for Nested Distributed Real-Time Database Systems (DRTDBS), emphasizing its enhancements in concurrency control and resource utilization. To further explore practical implementations, Section 7 presents a lock acquisition algorithm for nested transactions using Depth-First Search (DFS) and parallel Breadth-First Search (BFS). This section addresses three scenarios: no conflict, intra-transaction conflicts within the same tree, and inter-transaction conflicts across different trees, ensuring effective concurrency management. Continuing with performance analysis, Section 8 presents the experimental evaluation of Tree-HP2PL, covering research goals, MATLAB simulation setup, baseline protocol comparison, key performance metrics, and analysis of the protocol's efficiency and scalability. Looking ahead, Section 9 highlights Tree-HP2PL's strong performance and scalability across key metrics, noting trade-offs and future optimization needs. Section 10 highlights essential future directions for database management systems, emphasizing the need for advanced solutions, improved mechanisms, and intelligent optimization to meet the evolving demands of modern applications. Finally, Section 11 emphasizes the paper's key findings and suggests promising future research directions in nested transactions and concurrency control protocols, highlighting the importance of ongoing exploration in this evolving field to enhance efficiency and scalability in transaction management.

2 Related work

This section reviews the evolution of concurrency control mechanisms, particularly focusing on the challenges encountered in nested transactions within distributed real-time databases. It emphasizes the critical need for enhanced protocols to ensure global serializability and real-time performance. The discussion includes concurrency control protocols, highlighting their role in maintaining data consistency and addressing issues like deadlocks and resource contention. Additionally, it examines priority inversion in the Two-Phase Locking (2PL) protocol and explores the evolution of concurrency control and commit protocols tailored for nested transactions. This review underscores significant advancements while pointing out ongoing challenges, such as cascading intra-aborts, thereby establishing the necessity for specialized protocols to improve transaction processing efficiency in distributed environments.

In the realm of database management, the facilitation of concurrent access by multiple users is a critical aspect, fostering advantages such as heightened throughput, reduced transaction waiting times, and overall improved system performance. While read operations typically do not generate conflicts, the simultaneous execution of write operations introduces complexities leading to issues like

lost updates, uncommitted dependencies (or dirty reads), and inconsistent analyses. To counteract these challenges and ensure consistency and reliability in database operations, concurrency control mechanisms are indispensable. Although substantial progress has been made in real-time concurrency control for various transaction models, the distinctive challenges presented by nested transactions in distributed real-time databases require specialized protocols. The proposed work seeks to address these challenges by focusing on global serializability and real-time performance in the context of nested distributed transactions. This initiative contributes significantly to the evolving landscape of concurrency control protocols. Through a thorough examination of existing research, the related work provides a comprehensive overview of the historical evolution of concurrency control protocols and the development of commit protocols tailored for nested distributed real-time database systems. The proposed work aims to build upon these foundational aspects, offering innovative solutions to the specific challenges posed by nested transactions within the dynamic environment of distributed real-time databases. In essence, this research strives to bridge existing gaps, pushing the boundaries of knowledge, and contributing valuable insights to the intricate domain of concurrency control in distributed real-time databases.

To sum up, this research builds on the evolution of concurrency control mechanisms by addressing the unique challenges of nested transactions in distributed real-time databases. It offers innovative solutions to enhance global serializability and real-time performance, contributing valuable insights to the field and bridging critical gaps in existing protocols.

2.1 Concurrency control protocols

This section outlines the importance of concurrency control protocols in managing transactions in multi-user environments. It briefly discusses Pessimistic and Optimistic protocols, emphasizing their role in ensuring data integrity and consistency. The section also highlights challenges and solutions in real-time and distributed systems, stressing the need for robust concurrency control protocols.

Concurrency control is a critical aspect of database management systems to ensure transactions are executed in a controlled and consistent manner in multi-user environments. This involves managing access to shared resources to prevent conflicts and maintain data integrity. Serializability is a key concept, aiming to identify schedules where transactions can run concurrently without conflicts [14].This ensures that the database state mirrors a sequential execution. Two primary approaches to concurrency control are Pessimistic and Optimistic [33]. Pessimistic approaches, including lock-based algorithms, order-based algorithms, and hybrid approaches, assume conflicts are likely. Two-Phase Locking (2PL), assume conflicts are likely and address issues such as lost updates, uncommitted dependencies, and inconsistent analysis. Optimistic approaches, including optimistic locking [34] and timestamp ordering [35], [36], assume conflicts are infrequent and involve read, validation, and write phases. The Two-Phase Locking (2PL) protocol consists of growing and shrinking phases to ensure conflict serializability [37]. If every transaction strictly follows the Two-Phase Locking protocol in a schedule, the resulting schedule is guaranteed to demonstrate conflict serializability. Challenges associated with 2PL, such as cascading rollbacks, are addressed by rigorous 2PL and strict 2PL variants. The applicability of 2PL extends to distributed databases through variants such as Primary site 2PL [38], Primary Copy 2PL [39], Voting 2PL (Majority Consensus 2PL) [35], and Distributed Two-Phase Locking (D2PL), contributing to the mitigation of data redundancy. Practical applications of D2PL include System R* [40]and NonStop SQL [41], [42], [43]. Distributed Two-Phase Locking (D2PL) algorithms play a crucial role in minimizing data redundancy in distributed database systems. Timestamp Ordering (TO) is another approach where pessimistic TO prioritizes transactions based on timestamps to address conflicts. Basic TO ensures conflict serializability but introduces overhead, mitigated by Conservative TO algorithms [44]. Multiversion TO allows multiple versions of data items, enhancing concurrency [45], [46]. When a transaction reads data, the system chooses the version based on the transaction's timestamp, exploring it as an alternative to nested concurrency control protocols [47], [48], [49]. Conservative TO Algorithms introduce a delay to minimize transaction restarts and potential deadlock situations, with specific implementations providing solutions for reducing system overhead. Optimistic TO algorithms link timestamps exclusively to transactions, providing higher concurrency but facing challenges like increased storage overhead. Real-time database systems (RTDBS) face challenges such as restarts, useless restarts, and chained blocking [50]. Transaction priority schemes, including FCFS, EDF, MSTF, and SJF, aim to provide predictability in these systems [51], [52], [53].Understanding and implementing effective concurrency control protocols is crucial for maintaining the integrity and reliability of database systems. The diverse approaches outlined showcase the flexibility and adaptability of these protocols in addressing various challenges associated with concurrent transactions. As technology evolves, the choice of concurrency control becomes increasingly nuanced, with considerations for both pessimistic and optimistic scenarios. This overview serves as a starting point for delving into the fascinating realm of concurrency control, offering insights into the tools and techniques that underpin the foundation of robust and concurrent database systems.

Ultimately, the various concurrency control protocols discussed highlight their essential role in ensuring consistency and integrity in multi-user database environments. The choice between pessimistic and optimistic approaches, along with their adaptations in real-time and distributed systems, underscores the need for tailored solutions to address specific challenges. As technology evolves, these protocols will continue to play a crucial role in maintaining reliable and efficient transaction processing in modern database systems.

2.2 Priority inversion in 2pl (two-phase locking)

This section addresses priority inversion in Two-Phase Locking (2PL) within Real-Time Database Systems (RTDBS). It highlights the challenge of high-priority transactions being delayed by locks held by lower-priority transactions. Solutions, such as the Priority Abort Protocol and Priority Inheritance Protocol, are discussed to mitigate this issue and ensure timely execution of high-priority transactions.

In Real-Time Database Systems (RTDBS), priority inversion poses a significant challenge within the Two-Phase Locking (2PL) protocol. This occurs when high-priority transactions wait for locks held by lower-priority transactions, potentially causing delays and missed deadlines [54]. Two forms of priority inversion exist: bounded and unbounded [55]. To address this, solutions such as the Priority Abort Protocol (2PL-HP) involve aborting lower-priority lock holding transactions, allowing high-priority transactions to acquire them [4]. This solutions aim to ensure timely execution of high-priority transactions in RTDBS, though they may introduce certain drawbacks, such as increased restart frequency in the case of 2PL-HP.Another approach is the Priority Inheritance Protocol, which propagates priority from higher- to lower-priority transactions, preventing indefinite waits caused by lower-priority transactions holding required locks [56]. In summary, addressing priority inversion in 2PL within the context of RTDBS is crucial for ensuring the timely execution of high-priority transactions and meeting stringent real-time constraints. The outlined solutions, such as the Priority Abort Protocol and Priority Inheritance Protocol, provide avenues for tackling this challenge and maintaining the desired priority assignment goals in real-time database systems.

Overall, addressing priority inversion in Two-Phase Locking (2PL) within Real-Time Database Systems (RTDBS) is essential for ensuring that high-priority transactions meet their deadlines. The discussed solutions, including the Priority Abort Protocol and the Priority Inheritance Protocol, offer effective strategies to mitigate delays caused by lower-priority transactions, thereby enhancing the performance and reliability of RTDBS in real-time environments.

2.3 Evolution of concurrency control and commit protocols in nested transactions

This section explores the evolution of concurrency control and commit protocols for nested transactions. It presents an overview of various protocols in Table 1, highlighting their unique features, benefits, and challenges. The ongoing development of these protocols underscores the pursuit of scalability and efficiency, promising advancements in managing nested transactions. Database professionals are encouraged to consider these protocols for enhanced performance.

The landscape of database management has witnessed remarkable developments in concurrency control protocols, especially those tailored for nested transactions. This Table 1 presents a comprehensive overview of various protocols designed by different developers to address the unique challenges posed by nested transaction scenarios. Each protocol brings its own set of merits, demerits, and addresses specific challenges, contributing to the diverse toolkit available for database administrators. Let's delve into the details of these protocols, exploring their origins, distinctive features, and the challenges they aim to overcome. From fundamental algorithms to specialized approaches for distributed real-time systems, these protocols showcase the ingenuity of developers in the pursuit of efficient and robust nested transaction concurrency control.

Table 1: Overview of concurrency control and commit protocols for nested transactions.

S. No.	Protocol	Description	Advantages	Limitations	Improvement via Tree-HP2PL
1.	2PL-NT Concurrency control protocol [24]	Developed Concurrency Control Algorithm for Nested Transactions by integrating Eswaran's [37] two-phase locking mechanism.	Ensures serializability	High blocking and abort rates	DFS/BFS traversal reduces blocking and enhances concurrency

2.	Exclusive Locking Algorithm for Nested Transactions [57] [58]	Extending the multi-granularity algorithm tailored for nested transactions.	Precise access control	High locking overhead, reduced concurrency	Binary tree model reduces granularity complexity
3.	Formalization and Generalized Locking Algorithm [59]	Formal proof of Moss's read/write algorithm, introducing a generalized read-update locking algorithm.	Rigor and correctness	Complex to implement and scale	Simplifies lock modelling via structured hierarchy
4.	Extension of Multi-granularity Algorithms [60]	Extended multi-granularity algorithms for nested transactions.	Reduces lock table size	Reduced concurrency under escalated states	Avoids escalation, uses consistent hierarchical locking
5.	Serialization Graph Construction for Nested Transactions [61]	Serialization graph construction based on I/O automaton models, contributing to nested transaction systems.	Theoretical correctness	Computationally expensive in large-scale systems	Replaces graphs with traversal for efficiency
6.	Formalization of Concurrency Control Algorithms [62]	Formalizing concurrency control algorithms for open and safe nested transactions, utilizing an I/O approach.	Strong conflict resolution foundation	Lacks practical scalability	Practical DFS/BFS model supports real-time enforcement
7.	Application of Nested Transactions in KBMSs [63]	Applies nested transactions in knowledge base systems (KBMSs).	High abstraction flexibility	Inefficient in real-time scenarios	Supports real-time nested DBMS with structured control
8.	Extension of Gifford's basic quorum consensus algorithm for data replication to incorporate nested transactions and transactions aborts [64]	Extends Gifford's algorithm for data replication to include nested transactions.	Fault-tolerant consistency	High coordination overhead	Focuses on intra/inter conflict resolution without quorum delays
9.	Concurrency Control Algorithm for B-trees within nested transactions [65]	Proposed concurrency control algorithm specifically designed for B-trees within nested transaction models.	Optimized tree access for indexing	Limited to index concurrency	Tree-HP2PL generalizes to all data types, not just index structures
10.	Concurrency Control Algorithm utilizing linear hash	Presented a concurrency control algorithm utilizing linear hash structures for nested transactions.	Lightweight hashing model	Limited lock semantics for complex nesting	Extends to complex locking and transaction trees

	structures for nested transactions. [66]				
11.	Multi-version Timestamp Concurrency Control [45]	Introduces a multi-version timestamp concurrency control algorithm for nested transactions.	High read throughput	Storage overhead, version management	Avoids versioning, uses efficient priority-based lock conflict resolution
12.	2PL-NT-HP concurrency control protocol [67] [68]	Introduces concurrency control protocol for Nested Distributed Real-Time Database Systems	Deadline awareness	Still experiences priority inversion	Uses priority inheritance and abort handling for inversion-free scheduling
13.	S-PROMPT commit protocol [67] [68]	Specifically crafted commit protocol for nested transactions to address intra-abort cascade issues.	Ensures commit order [69].	Image storage overhead, added latency	Resolves conflicts early without image storage

The diverse array of concurrency control protocols for nested transactions showcased in the Table 1 highlights the ongoing efforts to refine and optimize database management in intricate scenarios. Each protocol addresses specific challenges while introducing new dimensions to the field. As developers continue to innovate and refine their approaches, the future promises even more sophisticated solutions for nested transaction concurrency control. The ongoing pursuit of scalability, compatibility, and efficiency in diverse database architectures will undoubtedly shape the next generation of protocols. This exploration serves as a snapshot of the current landscape, inviting database professionals to consider the nuances of each protocol in their quest for resilient and high-performance nested transaction management.

To sum up, the evolution of concurrency control and commit protocols for nested transactions highlights significant advancements in database management. The diverse protocols outlined in Table 1 showcase unique strategies to address challenges while offering distinct advantages. Ongoing innovations aim for scalability, compatibility, and efficiency, inviting database professionals to leverage these solutions for improved nested transaction management.

In summary, Section 2 analyses concurrency control mechanisms and commit protocols, revealing significant advancements in managing transaction complexities in distributed real-time databases. However, existing commit protocols frequently result in cascading intra-aborts within nested transactions, underscoring the urgent need for innovative solutions to enhance global serializability and real-time performance. This section emphasizes the necessity for further research into specialized protocols tailored for nested transactions, aiming to provide insights that contribute to efficient and reliable transaction processing in complex database environments.

3 Issues and challenges of nested transactions

This section examines the challenges of nested transactions in database management and proposes tailored solutions to address these complexities. It emphasizes the need for ongoing research to enhance performance and ensure data integrity in complex transactional environments.

The implementation of nested transactions introduces a unique set of challenges that demand innovative solutions for efficient database management. In the Table 2 below, we delve into specific issues encountered within the nested transaction model, along with their corresponding descriptions and proposed resolutions. Explore how developers and researchers have addressed challenges such as handling intra-transactions parallelism, prioritization policies for subtransactions, and the detection of deadlocks within nested transactions. Each issue is accompanied by thoughtful resolutions, ranging from concurrency control mechanisms to specialized protocols designed to preserve database coherence and ensure global serializability. This comprehensive overview sheds light on the complexities of managing nested transactions, providing insights into the ongoing efforts to enhance performance, maintain consistency, and overcome obstacles inherent in intricate transactional structures

Table 2: Nested transaction issues and remedies.

S. No	Issues in Nested Transaction Model	Description	Issues Resolution
I.	Handling of intra-transactions parallelism [32]	In nested transactions, both intra-transaction parallelism and inter-transaction parallelism are present. Intra-transaction parallelism in nested transaction models poses synchronization challenges during concurrent subtransaction execution, increasing the risk of deadlock and resource contention.	To address this issue, implement the following solutions: Concurrency Control Mechanisms, Isolation Levels, Transaction Scheduling Policies, Resource Management, Performance Monitoring, and Tuning, Documentation and Best Practices.
II.	Priority assignment policy for subtransactions [18]	Data sharing among subtransactions can cause delays, so prioritization is needed to avoid execution delays.	To address this issue, implement a priority assignment policy for subtransactions to ensure efficient execution and avoid delays.
III.	Detecting deadlock in nested transactions [70]	Nested transactions require both waits-for-lock and waits-for-commit relations for deadlock detection.	To address this issue, extend deadlock detection mechanisms to consider both waits-for-lock and waits-for-commit relations in the context of nested transactions.
IV.	Priority assignment policy [4] [71]	Traditional protocols for RTDBS often neglect transaction priorities.	To address this issue, develop and incorporate priority assignment policies in protocols for real-time database systems to consider transaction priorities.
V.	Concurrency control protocol for subtransactions [67], [68]	Existing concurrency control protocols designed for flat transactions may lead to issues when applied in a nested environment.	To address this issue, design concurrency control protocols specifically tailored for the concurrency of parent and child transactions in nested transaction models.
VI.	Commit protocol for subtransactions [67], [68]	Existing commit protocols may face challenges when applied to concurrent parent and child transactions in nested models.	To address this issue, develop commit protocols that address the unique challenges presented by concurrent execution of parent and child transactions in nested environments.
VII.	Handling of priority inversion [54]	Priority Inheritance and Priority Abort Protocols are used for priority inversion in advanced transaction models.	To address this issue, implement Priority Inheritance and Priority Abort Protocols to address priority inversion issues in nested transaction models.
VIII.	Preserving database coherence [28]	Nested transactions are not atomic; thus, the definition of "N-ACID" properties (nested-all-or-nothing, nested consistency, nested isolation, and nested durability) is necessary.	To address this issue, define and ensure N-ACID properties (N-A, N-C, N-I, N-D) for nested transactions to preserve database coherence.
IX.	Adjusting transaction recovery according to control structure [72]	Nested transactions have more splitable execution modules and finer control for recovery and concurrency compared to flat transactions.	To address this issue, adapt transaction recovery mechanisms to account for the specific control structure and finer granularity of recovery in nested transactions.
X.	Insurance of global serializability [67]	Nested transaction models have not been fully applied to real-time database systems, raising concerns about the global serializability of distributed real-time nested transactions.	To address this issue, research and develop methods to ensure the global serializability of distributed real-time nested transactions in real-time database systems.
XI.	Handling of transaction parameters in nested transaction [31].	The number of leaves and levels in nested transactions can impact performance.	To address this issue, consider and optimize transaction parameters, such as the number of leaves and levels, to enhance the system's performance in nested transaction models.

The Table 2 above encapsulates the multifaceted nature of challenges in nested transaction models and offers practical resolutions that span a spectrum of database management aspects. As we navigate the intricate landscape of nested transactions, it becomes evident that a tailored approach is crucial for ensuring seamless execution and maintaining data integrity. The ongoing research and development efforts highlighted in this exploration underscore the dynamic nature of database systems. By continually refining protocols, addressing deadlock scenarios, and introducing priority assignment policies, the database community strives to unlock the full potential of nested transactions. As technology evolves, so too will our ability to tackle the complexities of nested

transactions, fostering a future where robust and efficient database management becomes an integral part of intricate transactional scenarios.

In summary, the exploration of issues and challenges in nested transactions highlights the complexities inherent in managing such systems. By identifying key challenges and proposing targeted solutions, this section emphasizes the necessity for ongoing research and development to enhance performance and ensure data integrity. As the field evolves, the database community must continue to innovate and refine protocols to effectively navigate the intricacies of nested transactions, ultimately paving the way for more robust and efficient database management in complex environments.

4 Tree-based protocol for nested transactions

This section presents the Tree-Based Protocol for nested transactions, highlighting its hierarchical locking strategy, which enhances concurrency while ensuring conflict serializability and deadlock avoidance.

In the intricate realm of nested transactions, the choice of concurrency control protocols plays a pivotal role in shaping the efficiency and reliability of database systems. The Tree-Based Protocol for concurrency control in nested transactions, employs Exclusive Locks Only with a Parent-Child Relationship structure. Locks can be released during the transaction, promoting high concurrency. The approach assures Conflict Serializability, prevents deadlocks hierarchically, and reduces waiting times. Despite potential Locking Overhead, careful optimization is required. Overall, the protocol offers adaptability and effectiveness in various nested transaction scenarios with critical concerns like conflict serializability and deadlock prevention.

Thus, the Tree-Based Protocol for nested transactions enhances concurrency while ensuring conflict serializability and deadlock prevention. Its hierarchical locking mechanism promotes efficiency, though careful optimization is needed to manage potential locking overhead. Overall, it offers a strong framework for improving nested transaction management in complex database systems.

5 Tree-HP2PL: A tree-based concurrency control protocol for nested DRTDBS

This section explores the Tree-HP2PL protocol, a concurrency control mechanism designed to optimize nested transaction management. The Table 3 delves into key aspects of this protocol, highlighting its innovative features and advantages in enhancing system performance and conflict resolution, making it suitable for the complexities of nested transactions.

In the realm of nested transactions, the choice of concurrency control mechanisms profoundly impacts system efficiency and reliability. The Table 3 delves into key aspects of a sophisticated concurrency control protocol designed to optimize nested transaction management. TreeHP2PL protocol leverages a compatible tree data structure, employing a binary search tree for streamlined conflict resolution and efficient traversal. It addresses both intra-transaction and inter-transaction conflicts through innovative priority inheritance mechanisms, while strategically mitigating pseudo priority inversion. Explore the nuances of this protocol, from its utilization of graph traversal algorithms for lock-holding transaction identification to the advantages it brings in terms of performance optimization and simplified conflict resolution. Each aspect contributes to the overall efficiency of the system, making it well-suited for the intricacies of nested transactions. Tree-HP2PL ensures conflict serializability by acquiring locks using structured DFS and parallel BFS traversal, which prevents cycles and preserves the correct ordering of conflicting operations. Conflict serializability refers to the ability to reorder non-conflicting operations in a concurrent schedule to match a valid serial execution. Let's dissect the protocol's characteristics, highlighting its key points, and understanding how it contributes to the optimization of nested transaction management.

Table 3: Key Aspects of the TREEHP2PL in nested transactions

S.No.	Aspect	Description	Key Points
1.	Compatible Tree Data Structure	Utilizes a binary search tree for optimized system performance, enabling straightforward in-order traversal for conflict resolution.	Performance Optimization: Compatible tree structure enhances efficiency over array-based approaches.
2.	Intra-Transaction Data Conflict Handling	Employs binary search tree traversal method to streamline conflict resolution, avoiding multiple insertions into the waitlist of data items.	Simplified Conflict Resolution: Binary search tree traversal simplifies intra-transaction conflict handling.
3.	Priority Inheritance Mechanism	Upgrades priorities of nodes in a subtree for incoming high-priority sub-	Effective Priority Inheritance: Ensures effective resolution of intra- and inter-transaction

		transactions, minimizing extensive priority comparisons and enhancing efficiency.	conflicts through priority inheritance.
4.	Inter-Transaction Data Conflict Resolution	Uses priority inheritance and priority abort protocols at the parent transaction level to address conflicts, prioritizing nested transactions.	Effective Resolution of Inter-Transaction Conflicts: Priority inheritance and priority abort protocols employed for conflict resolution.
5.	Handling Pseudo Priority Inversion	Strategically employs priority inheritance and priority abort protocols, with a preference for priority inheritance in nested transactions to address pseudo priority inversion.	Priority Inversion Mitigation: Addresses pseudo priority inversion, enhancing overall system performance.
6.	Lock-Holding Transaction Identification	Employs Depth-First Search (DFS) and parallel Breadth-First Search (BFS) as independent yet complementary graph traversal methods. DFS enables deep dependency exploration, while parallel BFS supports concurrent conflict resolution in wider transaction hierarchies.	Effective Lock-Holding Transaction Identification: Graph traversal algorithms assist in identifying lock holding transactions
7.	Advantages and Contributions	- Performance Optimization: Adoption of a compatible tree structure enhances efficiency over array-based approaches. - Simplified Conflict Resolution: Binary search tree traversal simplifies intra-transaction conflict handling. - Effective Priority Inheritance: Ensures effective resolution of intra- and inter-transaction conflicts through priority inheritance. - Priority Inversion Mitigation: Addresses pseudo priority inversion, enhancing overall system performance.	Performance benefits through optimized data structure. Streamlined conflict resolution within and between transactions. Efficient handling of priority inheritance and pseudo priority inversion.

The detailed exploration of the optimized concurrency control protocol for nested transactions reveals a sophisticated approach to addressing conflicts and enhancing system performance. As we step beyond the Table 3, it's crucial to consider the broader implications and potential applications of the protocol. The advantages and contributions highlighted, ranging from performance optimization to effective resolution of inter-transaction conflicts, underscore the protocol's versatility in diverse nested transaction scenarios. The strategic use of priority inheritance mechanisms and graph traversal algorithms for lock-holding transaction identification showcases a holistic approach to concurrency control. Database administrators and developers navigating complex transactional environments are encouraged to assess how these nuanced features align with their specific requirements. Considerations extend beyond the technical aspects, touching upon scalability, adaptability, and ease of integration within existing database architectures. This exploration invites further scrutiny, experimentation, and adaptation as the database community strives to optimize nested transaction management. Beyond the structured presentation, there lies a landscape of possibilities for refinement and customization based on unique system demands and evolving technology.

In summary, the Tree-HP2PL protocol presents a robust and innovative approach to managing concurrency in nested transactions, significantly enhancing system performance and conflict resolution. By utilizing a compatible tree structure and incorporating advanced mechanisms such as priority inheritance and graph traversal algorithms, the protocol effectively addresses both intra- and inter-transaction conflicts. The insights from Table 3 reinforce its potential to optimize nested transaction management, making it a valuable consideration for database administrators and developers facing the challenges of complex transactional environments. Overall, this exploration underscores the importance of selecting appropriate concurrency control mechanisms to ensure the reliability and efficiency of nested transactions.

6 TREE-HP2PL model

This section outlines the experimental evaluation of the Tree-HP2PL protocol for nested transactions. It begins with the research objectives and simulation setup, detailing the use of MATLAB to model varying transaction volumes and conflict scenarios. The next part introduces the baseline protocol (2PLNTHP) and the evaluation criteria used for comparison. Key performance metrics—such as

execution time, response time, memory usage, system throughput, and success ratio—are then defined. The final part provides a detailed analysis of these metrics to assess the efficiency, scalability, and reliability of the Tree-HP2PL protocol in real-time distributed environments.

The TreeHP2PL model, designed for Nested Distributed Real-Time Database Systems (DRTDBS), comprises several key components that collectively contribute to its effectiveness in managing concurrency control. The system consists of four main components: the Source Module, responsible for simulating external workload; the Transaction Manager, handling transactions and assigning priorities; the Concurrency Control Manager, managing conflicts and coordinating transactions; and the Resource Manager, controlling physical resources for efficient allocation. Together, they ensure effective transaction management, modelling, and resource utilization within the computing environment.

Ultimately, the TreeHP2PL model significantly enhances the management of concurrency in Nested Distributed Real-Time Database Systems. By integrating its core components, the model optimizes transaction management and resource utilization, ensuring robust performance in complex distributed environments. This comprehensive approach positions the model as a valuable solution for addressing the challenges of concurrency control in real-time applications.

7 Proposed algorithm

This section introduces a lock acquisition algorithm for nested transactions, focusing on conflict management at various levels. It employs Depth-First Search (DFS) and parallel Breadth-First Search (BFS) to enhance scalability and conflict resolution. The algorithm is structured to handle three key conflict scenarios: no conflict, intra-transaction conflicts within the same transaction tree, and inter-transaction conflicts among different transaction trees. The algorithm emphasizes efficiency and adaptability through a systematic approach, which includes initialization, lock checks, and conflict resolution steps. By employing nested functions, the algorithm ensures a thorough mechanism for lock acquisition that can be evaluated and customized for practical implementation in real-world database systems. The combination of DFS and parallel BFS strategies demonstrates a commitment to optimizing transaction management within nested structures, encouraging further exploration and refinement by database professionals.

7.1 Lock acquisition

Effective lock acquisition is a critical aspect of nested transaction management, ensuring both concurrency and data integrity. The proposed algorithm outlined below provides a comprehensive strategy for handling various conflict scenarios within nested transactions.

This algorithm addresses three key cases:

1) No Conflict Among Transactions

In the absence of conflicts, the algorithm grants locks to transactions that have not yet acquired them.

2) Intra-Transactions Conflict

When conflicts arise within the same transaction tree, the algorithm employs a sophisticated conflict resolution mechanism utilizing Depth-First Search (DFS) and parallel Breadth-First Search (BFS) strategies.

3) Inter-Transactions Conflict

In situations where conflicts occur among different transaction trees, the algorithm leverages the same conflict resolution mechanism used for intra-transaction conflicts. The algorithm employs nested functions, utilizing DFS and parallel BFS to resolve conflicts efficiently. It initializes data structures, transaction trees, priorities, and shared data to facilitate a systematic approach to lock acquisition. As we delve into the algorithm's intricacies, it becomes apparent that the proposed approach is not only geared towards resolving conflicts but is also designed for scalability and adaptability within diverse nested transaction scenarios. Let's explore the detailed steps and mechanisms embedded in the algorithm, shedding light on its capacity to optimize lock acquisition and contribute to the overall efficiency of nested transaction systems.

Here are the step-by-step instructions for the lock acquisition algorithm.

1. Step 1: Initialize Global Variables:
   - Create a set named visited to keep track of visited transactions.
   - Create a dictionary named transactionTree to represent the transaction tree.
   - Create a list named transactionPriorities to store transactions prioritized for conflict resolution.
   - Create a dictionary named sharedData to store the status of shared data.
2. Step 2: Check if any Transaction has not Acquired a Lock:
   - Implement A Function or check condition to verify if any transaction has not acquired a lock.
3. Step 3: Grant a Lock to a Transaction:

- Implement a function or set of conditions to grant a lock to a transaction if it has not acquired one.

4. Step 4: Handle Intra-Transaction Conflict Resolution:
   - Initialize necessary data structures and variables.
   - Initialize the transaction tree and transaction priorities.
   - Implement a depth-first search (DFS) and parallel breadth-first search (BFS) combination for resolving intra-transaction conflicts within the same transaction tree.
   - Prioritize transactions based on a specified order for resolution.
5. Step 5: Handle Inter-Transaction Conflict:
   - Defer to the mechanism used for intra-transaction conflict resolution to handle conflicts among different transaction trees.
6. Step 6: Main Algorithm Execution:
   - Check if any transaction has not acquired a lock and grant the lock if necessary.
   - Handle intra-transaction conflicts using the previously defined mechanism.
   - Handle inter-transaction conflicts by deferring to the intra-transaction conflict resolution mechanism.

These steps provide a high-level overview of the lock acquisition algorithm.

7.2 Algorithm

The lock acquisition algorithm adopts a systematic and adaptive approach to manage conflicts within nested transactions. Its versatility is highlighted through its ability to address conflicts at different levels within the transaction hierarchy.

Input:

T: Set of nested transactions

transactionTree: Hierarchical structure of parent-child transactions

sharedData: Lock status for each data item

visited: Set to keep track of visited transactions

Output:

Locks granted or conflicts resolved

BEGIN

CASE 1: No Conflict Among Transactions

For each transaction Ti ∈ T do

If Ti has not acquired a lock and sharedData[Ti.dataItem] is free then

grant_lock(Ti)

End If

End For

CASE 2: Intra-Transaction Conflict (Same Tree)

Procedure handle_intra_transaction_conflict()

initialize_data_structures()

initialize_transaction_tree()

initialize_transaction_priorities()

For each transaction in transactionPriorities do

initialize_shared_data(transaction)

End For

Start conflict resolution:

start_DFS_or_parallel_BFS(root_node, selected_transaction, sharedData)

Function DFS(transactionIndex, sharedData)

resolve_conflict(transactionIndex, sharedData)

Function resolve_conflict(transactionIndex, sharedData)

If transactionIndex ∉ visited then

visited ← visited ∪ {transactionIndex}

process_transaction(transactionIndex)

For each child ∈ transactionTree[transactionIndex] do

```
If child is unresolved then
DFS(child, sharedData)
Else
resolve_intra_transaction_conflict(transactionIndex, child)
End If
End For
End If

Function parallel_BFS(sharedData)
queue ← ∅
initialize_data_structures()
While queue ≠ ∅ do
transaction ← dequeue(queue)
For each child ∈ transactionTree[transaction] do
If child is unresolved then
process_transaction(child)
Else
resolve_intra_transaction_conflict(transaction, child)
End If
enqueue_unvisited_children(queue, transaction)
    End For
    End While

CASE 3: Inter-Transaction Conflict (Different Trees)
Procedure handle_inter_transaction_conflict()
Call handle_intra_transaction_conflict()

END
```

The intricacies of the proposed lock acquisition algorithm unveil a systematic and adaptive approach to managing conflicts within nested transactions. As we move beyond the algorithm, it's essential to reflect on the broader implications and considerations for implementation. The algorithm's versatility shines through its ability to handle conflicts at different levels, ensuring that transactions progress smoothly even in complex nested structures. The combination of DFS and parallel BFS strategies showcases a commitment to efficiency and scalability, crucial for real-world applications. Database administrators and developers are encouraged to evaluate how this algorithm aligns with the specific demands of their nested transaction scenarios. Considerations extend beyond the algorithmic steps, touching upon factors like system performance, adaptability to varying transaction loads, and ease of integration into existing databases. This exploration invites further examination, experimentation, and customization as the database community strives to optimize lock acquisition in the intricate landscape of nested transactions. Beyond the structured presentation of the algorithm, lies a realm of possibilities for refinement, tailoring the approach to unique system requirements and evolving technological landscapes.

In summary, the proposed lock acquisition algorithm for nested transactions offers a robust framework for managing conflicts through a systematic approach that integrates Depth-First Search (DFS) and parallel Breadth-First Search (BFS) strategies. By effectively addressing no conflicts, intra-transaction conflicts, and inter-transaction conflicts, the algorithm ensures efficient lock management while maintaining data integrity. Its adaptability to various transaction scenarios highlights its potential for real-world application, encouraging database professionals to customize and refine the algorithm to meet specific system needs. This exploration serves as a foundation for further research and development in optimizing transaction management within complex nested structures.

8 Implementation and results

This section outlines the experimental evaluation of the Tree-HP2PL protocol for nested transactions. It begins with the research objectives and simulation setup, detailing the use of MATLAB to model varying transaction volumes and conflict scenarios. The next part introduces the baseline protocol (2PLNTHP), and the evaluation criteria used for comparison. Key performance metrics—such as execution time, response time, memory usage, system throughput, and success ratio—are then defined. The final part provides a detailed analysis of these metrics to assess the efficiency, scalability, and reliability of the Tree-HP2PL protocol in real-time distributed environments.

Tree-HP2PL algorithm's performance in nested transactions is assessed focusing on system throughput, success ratio, response time, execution time, and memory usage. These metrics offer deep insights into protocol performance and efficiency in handling concurrent transactions. Specifically for tree-based protocols such as Tree-HP2PL in nested transactions, focusing on these metrics enables a comprehensive evaluation of their real-world applicability and effectiveness. To validate our approach, we implemented the algorithm in a MATLAB simulation environment and conducted extensive

experiments using various transaction scenarios [73]. The study investigates the suitability of Tree-HP2PL in nested transaction scenarios, employing the Breadth-First Search (BFS) algorithm for tree and graph data structure traversal and search operations. Additionally, this protocol utilizes Depth-First Search (DFS) and parallel Breadth-First Search (BFS) strategies to effectively manage conflicts and improve scalability. Evaluating system throughput helps in understanding the protocol's ability to process transactions efficiently within a given timeframe. Success ratio indicates transaction completion rates and reliability. Response time measures system responsiveness to transaction requests, impacting user experience. Execution time reflects transaction processing efficiency and resource utilization. Assessing memory usage is crucial for scalability and resource management. Table 4 summarizes these performance metrics, providing a quick reference. The subsequent sections delve into a detailed analysis of Tree-HP2PL's average system throughput, success ratio, response time, execution time, and memory usage under varying transaction workloads. The protocol is scrutinized as it processes varying numbers of transactions (5, 10, 15, 20, 25, and 30) over ten iterations, simulating nested transactions in each iteration. Table 4 summarizes the analysis of average execution time trends across different transaction numbers over ten iterations, while Table 6 presents a similar analysis of average memory consumption. Tables 7 to 9 depict the trends in average system throughput, success ratio, and response time under varying transaction volumes and iterations. Analyzing the nested transaction performance within the experimental transaction tree involves a detailed examination of the shapes illustrated in Figures 1–6. Memory-resident and disk-resident databases represent different approaches to database storage, each with its own set of considerations. In the case of a memory-resident database, the entire database is assumed to be stored in the main memory. This simplifies the study as it eliminates the need to model disks or manage I/O scheduling. On the other hand, in a disk-resident database, we assume a more realistic scenario where the database is stored on secondary storage (disks). Disk-resident database approach is chosen for our study because it provides a practical representation of how data is typically stored and accessed. Additionally, the database may be on a single disk or partitioned across multiple disks, with each disk having its own service queue.

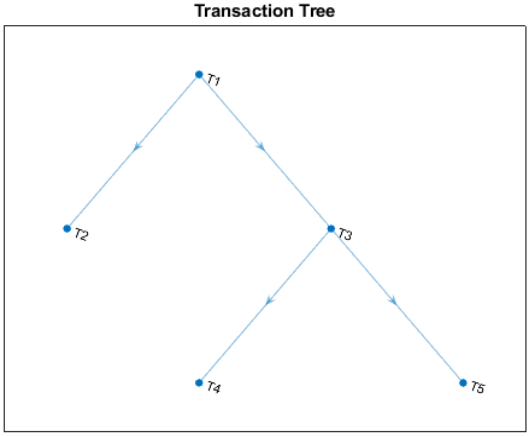


Figure 1: Five node transaction structure.

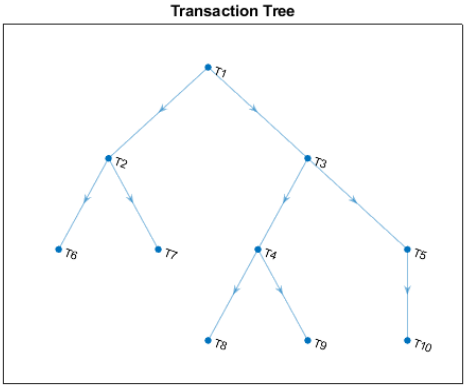


Figure 2: Ten node transaction structure.

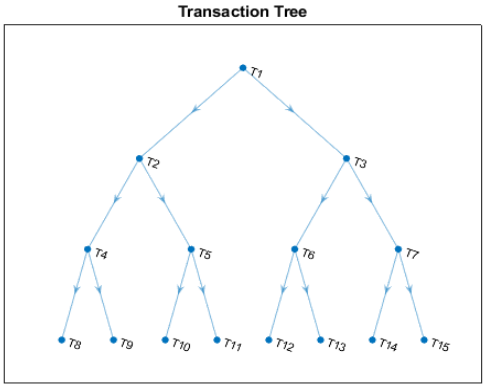


Figure 3: Fifteen node transaction structure.

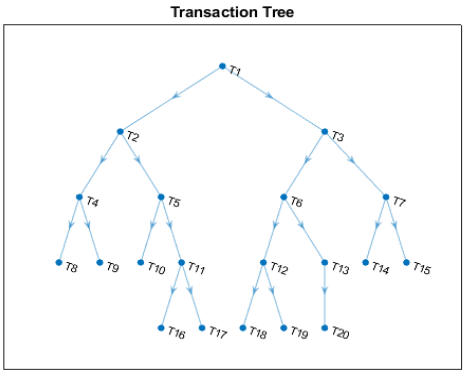


Figure 4: Twenty node transaction structure

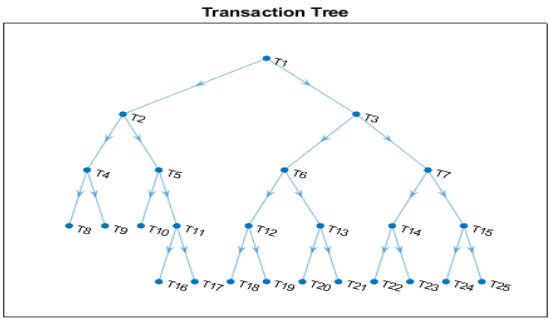


Figure 5: Twenty-five node transaction structure.

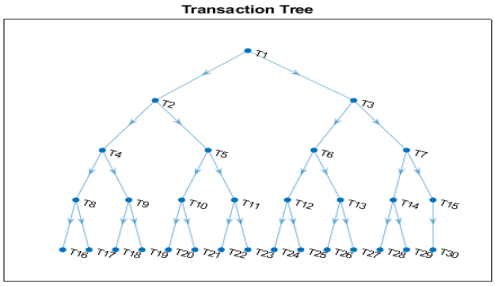


Figure 6: Thirty node transaction structure.

Overall, this section underscores the significance of the Tree-HP2PL algorithm within tree-based concurrency control protocols for nested transactions. It emphasizes the need to analyze various performance metrics in both memory-resident and disk-resident database contexts, thereby establishing a practical framework for understanding how these protocols function in real-world scenarios. The insights gained pave the way for future research to enhance the efficiency and applicability of concurrency control methods in nested transactions.

8.1 Research objectives and simulation setup

This section outlines the goals of evaluating Tree-HP2PL and describes the simulation environment used to assess its performance under nested transaction workloads.

This study aims to evaluate the effectiveness of the proposed Tree-HP2PL protocol for managing nested transactions in distributed real-time database systems (DRTDBS). The research is guided by the following questions☹1) How does Tree-HP2PL impact system throughput, execution time, memory usage, and response time across varying transaction volumes?(2) Can a hybrid Depth-First Search (DFS) and parallel Breadth-First Search (BFS) lock acquisition strategy enhance conflict detection and resolution in nested transactions?(3) What trade-offs arise from using hierarchical priority inheritance in real-time concurrency control? To answer these questions, Tree-HP2PL was implemented and evaluated in MATLAB R2023b on a system with an Intel Core i7-11800H processor (2.3 GHz), 16 GB RAM, and Windows 11 Pro (64-bit). The simulation modelled nested transaction workloads as binary trees with up to four levels of nesting. In each iteration, 5 to 30 transactions were executed, across 10 iterations. Conflict scenarios were introduced using random transaction priorities and shared dependencies to simulate intra- and inter-transaction contention. The workload abstracts real-world scheduling patterns but excludes specific database queries or logic. The five measured performance metrics include success ratio, system throughput, average execution time, average response time, and memory usage. The simulation assumes firm deadlines, a centralized lock manager with no network delays, and failure due to lock denial or timeout only, thus isolating concurrency control behavior under realistic DRTDBS-like constraints.

8.2 Baseline protocols and evaluation criteria

This section compares the proposed Tree-HP2PL protocol with the 2PLNTHP protocol using key performance metrics, based on analytical evaluation and literature data.

To evaluate the effectiveness of the proposed Tree-HP2PL protocol, its performance was compared analytically with one of the established concurrency control protocols for nested transactions: 2PLNTHP (Two-Phase Locking for Nested Transactions with High Priority), which incorporates transaction priorities into the traditional two-phase locking mechanism to improve deadline adherence. This protocol was selected due to their relevance to real-time nested transaction environments in distributed database systems. Since source code or implementation models for this protocol is not publicly available, direct benchmarking was not feasible. Therefore, the comparative analysis is based on performance trends reported in prior literature [67], [68] and is supported by analytical reasoning based on their documented operational characteristics. The evaluation considered five key performance metrics: average execution time, average response time, system throughput

(transactions per second), memory usage, and success ratio. While Tree-HP2PL was implemented and tested in MATLAB, data for the baseline protocols were derived from related studies under comparable transaction volumes and workload conditions, with the aim of identifying performance patterns rather than providing exact numerical equivalence.

8.3 Performance metrics

This section discusses key performance metrics for evaluating tree-based concurrency control protocols in nested transactions. In the evaluation of concurrency control protocols for transactions, various performance metrics are employed to assess the effectiveness and efficiency of the system [71]. These metrics provide a comprehensive understanding of different aspects of performance, aiding stakeholders in making informed decisions and implementing improvements. Table 4 summarizes the key performance metrics used to evaluate concurrency control protocols for transactions. These metrics play a crucial role in assessing the efficiency, success rate, and overall performance capacity of the system under evaluation.

Table 4: Performance metrics for concurrency control protocols

S.No.	Performance Metric	Description
1.	Average System Throughput	Measures the average number of transactions processed per unit of time, indicating system capacity.
2.	Average Success Ratio	Average ratio of successful transactions to total transactions submitted, showing system success rate.
3.	Average Response Time	Average elapsed time from command initiation to completion, assessing transaction performance.
4.	Average Execution Time	Average typical duration for completing each transaction, reflecting system efficiency.
5.	Average Memory Usage	Average amount of memory utilized per transaction, indicating system memory efficiency.

In summary, the evaluation of tree-based concurrency control protocols for nested transactions, specifically using the Tree-HP2PL algorithm, highlights the importance of key performance metrics. The analysis of these metrics provides valuable insights into the system's efficiency in handling nested transactions. This research underscores the effectiveness of BFS in managing transaction complexity and provides a foundation for further exploration of concurrency control mechanisms.

8.4 Performance metrics analysis

This section explores key performance metrics—System Throughput, Success Ratio, Response Time, Execution Time, and Memory Usage to assess system efficiency and optimize concurrency control protocols.

Before delving into the analysis of these performance metrics, it's essential to highlight their significance in evaluating concurrency control protocols. These metrics, including Average System Throughput, Average Success Ratio, Average Response Time, Average Execution Time, and Average Memory Usage, provide a comprehensive view of the system's efficiency, success rate, transaction performance, and resource utilization.

Understanding these metrics is key to making informed decisions and implementing improvements in concurrency control protocols for transactions. The following section will delve into a detailed analysis of each metric, shedding light on their impact and implications for system performance.

Thus, these metrics are crucial for optimizing concurrency control protocols by improving transaction management, resource utilization, and overall system performance.

8.4.1 Analysis of execution time

This section analyses average execution time across varying transaction volumes, focusing on the system's processing efficiency. It highlights trends and fluctuations in execution times, providing insights into system behavior under different workloads.

Table 5: Execution time for simulation.

Number of Transactions	Average Execution Time (seconds)
5	0.001125305
10	0.001727616
15	0.002153896
20	0.004914052
25	0.00276094
30	0.004611946

The analysis of average execution time provides crucial insights into the system's processing efficiency under varying transaction workloads. Table 5 summarizes the recorded average execution times for different

numbers of transactions over ten iterations, offering a detailed view of how execution time scales with transaction volume. This data is visually presented in Figure 7, showcasing the execution time trends, and highlighting any notable patterns or outliers across the transaction workload spectrum.

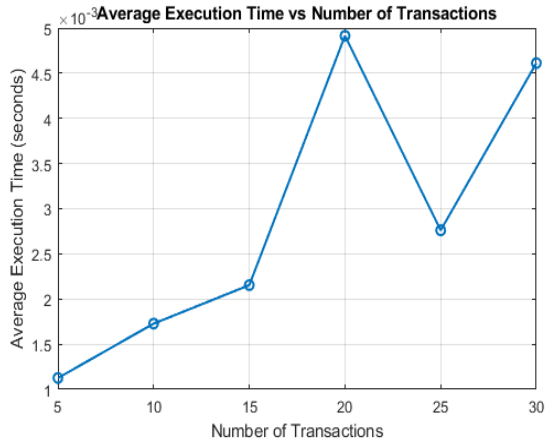


Figure 7: Execution time for simulation.

The analysis of average execution time reveals efficient processing at lower transaction volumes, with execution times as follows: 0.00799869 seconds for 5 transactions, 0.01193914 seconds for 10 transactions, and 0.01595354 seconds for 15 transactions. However, a significant increase is observed at 20 transactions, with execution time reaching 0.03583039 seconds, indicating potential strain or bottlenecks under heavier workloads. The execution time then slightly decreases for 25 transactions (0.02059183 seconds) and increases again for 30 transactions (0.03478573 seconds).

Overall, the analysis indicates efficient processing at lower transaction volumes, while a significant increase at higher volumes suggests potential bottlenecks, highlighting the need for optimizations to improve system performance under heavier workloads.

8.4.2 Analysis of memory usage

This section analyses average memory usage across varying transaction volumes, highlighting trends and the need for efficient memory management strategies to address increasing resource demands.

Table 6: Memory usage for simulation

Number of Transactions	Average Memory Usage (MB)
5	0.006073418
10	0.007920208
15	0.009810867
20	0.011739674
25	0.013970184
30	0.015982914

The analysis of average memory usage provides valuable insights into the system's memory, requirements under varying transaction workloads. Table 6 presents the recorded average memory usage values for different numbers of transactions over ten iterations, illustrating how memory usage scales with transaction volume. This data is visually represented in Figure 8, offering a clear depiction of memory usage trends, and highlighting any notable patterns or fluctuations across different transaction volumes.

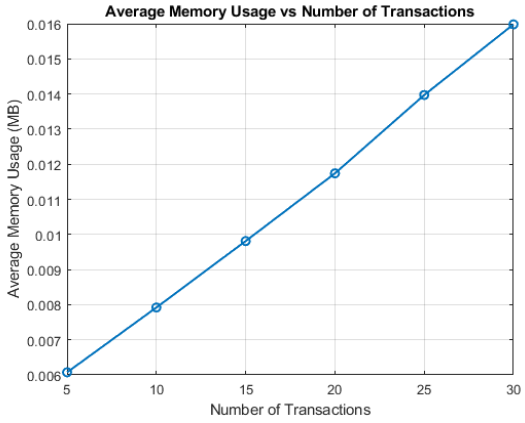


Figure 8: Memory usage for simulation.

Memory usage exhibits a linear growth pattern, increasing from 0.006073418 MB for 5 transactions to 0.015982914 MB for 30 transactions. This consistent increase emphasizes the importance of efficient memory management strategies to handle escalating resource demands effectively.

In summary, the analysis reveals a linear increase in memory usage with higher transaction volumes, underscoring the need for efficient memory management strategies to accommodate escalating resource demands and maintain optimal system performance.

8.4.3 Analysis of system throughput

This section examines average system throughput to evaluate processing capacity under varying transaction workloads, highlighting trends and emphasizing the need for optimizations to address potential bottlenecks.

The assessment of average system throughput provides essential insights into the system's processing capacity under varying transaction workloads. Table 7 summarizes the average throughput values recorded for different numbers of transactions across ten iterations, showcasing how throughput changes with transaction volume. This information is visually depicted in Figure 9, offering a graphical representation of throughput trends, and highlighting any significant fluctuations or trends observed across the range of transaction volumes.

Table 7: System throughput for simulation

Number of Transactions	Average System Throughput (TPS)
5	1980.904429
10	1563.57411
15	1235.399642
20	721.2252535
25	1374.468946
30	998.2999971

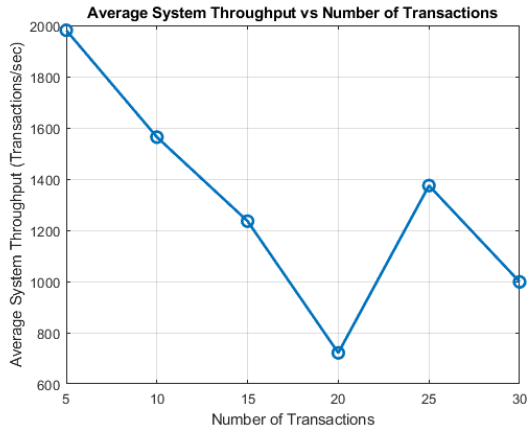


Figure 9: System Throughput for simulation.

System throughput shows a decline as transaction volumes increase, dropping notably from 1980.904429 TPS at 5 transactions to 721.2252535 TPS at 20 transactions. This decline suggests potential processing constraints or bottlenecks under higher transaction volumes, highlighting the need for optimizations to maintain optimal processing capacity.

Thus, the analysis indicates a significant decline in average system throughput with increasing transaction volumes, highlighting the need for optimizations to address potential bottlenecks and maintain optimal processing capacity.

8.4.4 Analysis of success ratio

This section evaluates the average success ratio to assess the system's reliability and transaction completion rates across varying workloads, highlighting the system's effectiveness in handling transactions.

The evaluation of average success ratio offers critical insights into the system's reliability and transaction completion rates across varying transaction workloads. Table 8 compiles the average success ratio values recorded for different numbers of transactions over ten iterations, providing a comprehensive view of success rates across transaction volumes. This data is visually represented in Figure 10, presenting success ratio trends graphically and highlighting any notable patterns or variations observed in the system's performance regarding transaction completions.

Table 8: Success ratio for simulation

Number of Transactions	Average Success Ratio
5	1
10	1
15	1
20	1
25	1
30	1

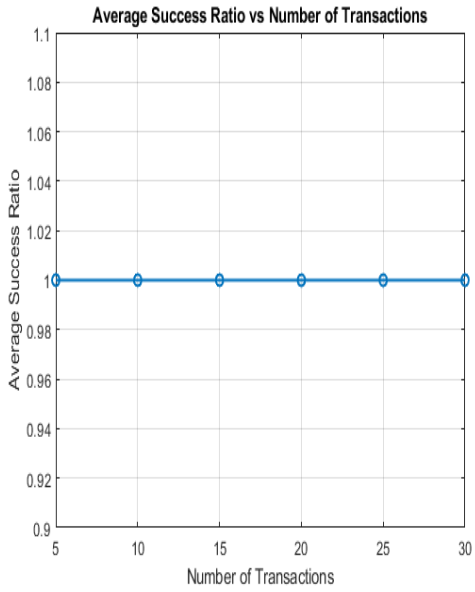


Figure 10: Success ratio for simulation.

The success ratio remains consistently high at 1 (or 100%) across all transaction volumes, indicating reliable transaction completion rates without errors or failures.

In summary, the analysis of the average success ratio underscores the system's reliability and effectiveness in managing transactions across varying workloads, indicating a robust performance in achieving transaction completions without errors.

8.4.5 Analysis of response time

This section analyses average response time, highlighting fluctuations in system responsiveness as transaction workloads change. It emphasizes the correlation between transaction volume and response time, suggesting implications for processing efficiency and resource management.

The assessment of average response time reveals insights into the system's responsiveness under varying transaction workloads. Table 9 summarizes the recorded average response time values for different numbers of transactions over ten iterations, illustrating how response time fluctuates with transaction volume. This information is graphically depicted in Figure 11, providing a visual

representation of response time trends, and highlighting any significant changes or patterns observed in the system's responsiveness across different transaction volumes.

Table 9:Response time for simulation.

Number of Transactions	Average Response Time (seconds)
5	0.00799869
10	0.01193914
15	0.01595354
20	0.03583039
25	0.02059183
30	0.03478573

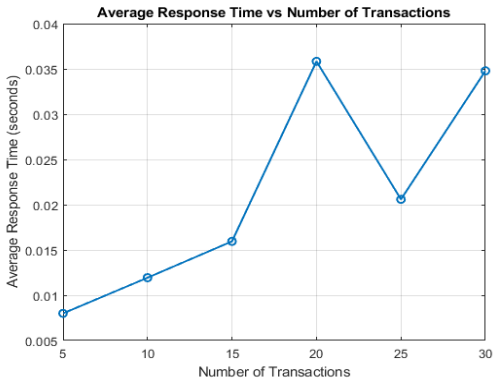


Figure 11: Response time for simulation.

Response time follows a pattern like execution time, with lower values at lower transaction volumes and increased times at higher volumes. Specifically, response times are 0.00799869 seconds for 5 transactions, 0.01193914 seconds for 10 transactions, and 0.01595354 seconds for 15 transactions. A notable increase is observed at 20 transactions (0.03583039 seconds), suggesting potential processing complexities or resource constraints impacting system responsiveness. The response time then decreases slightly for 25 transactions (0.02059183 seconds) and increases again for 30 transactions (0.03478573 seconds).

Thus, the analysis of average response time reveals that system responsiveness declines as transaction volumes increase, indicating potential processing challenges. Effective resource management is essential to maintain optimal performance under varying workloads.

Overall, the in-depth analysis of the system's performance across different transaction volumes revealed significant patterns and trends. The findings highlighted the effectiveness of the Tree-HP2PL algorithm, leveraging Depth-First Search (DFS) and parallel Breadth-First Search (BFS) strategies, in managing nested transactions and complex structures. The BFS approach demonstrated positive correlations among key metrics—average system throughput, success ratio, response time, execution time, and memory usage—relative to transaction volume. The analysis emphasized the critical role of workload management strategies, showcasing efficient processing at lower volumes with consistently high success rates, indicative of robust reliability. However, increased workloads led to notable spikes in execution and response times, suggesting potential bottlenecks impacting responsiveness. The linear growth in memory usage underscored the necessity for effective memory management, while the decline in system throughput with higher workloads highlighted the need for optimizations to maintain processing capacity. Ultimately, enhancing resource utilization and system responsiveness are imperative for sustaining performance across varying workloads.

9 Discussion

This section discusses the Tree-HP2PL protocol's strong performance and scalability based on five key metrics—success ratio, system throughput, execution time, response time, and memory usage—while noting trade-offs and opportunities for future real-world optimization.

The Tree-HP2PL protocol was evaluated through MATLAB simulation environment across transaction volumes ranging from 5 to 30, focusing on five performance metrics: success ratio, system throughput, average execution time, average response time, and memory usage. The protocol consistently achieved a 100% success ratio, demonstrating robust conflict resolution without deadlocks or aborts across all tested scenarios. System throughput declined gradually from approximately 1980 transactions per second (TPS) at 5 transactions to about 998 TPS at 30 transactions, reflecting controlled degradation under increasing load. Average execution time increased from 0.0011 seconds at 5 transactions to 0.0049 seconds at 20 transactions, while average response time remained below 0.017 seconds up to 15 transactions but rose to 0.036 seconds as contention and nesting depth increased. Memory usage exhibited a linear growth trend, from 0.006 MB to 0.016 MB, in line with the number of active transactions and the size of transaction tree structures maintained during execution.

Upon deeper analysis, the sharp rise in execution time beyond 20 transactions is primarily attributed to the algorithmic complexity inherent in Tree-HP2PL's conflict resolution strategy, which employs Depth-First Search (DFS) and parallel Breadth-First Search (BFS). The worst-case time complexity of both DFS and BFS components is O (V + E), where V is the number of transactions and E is the number of transaction dependencies. As transaction volume increases, so does the traversal workload, with more frequent updates to inherited priorities and increased BFS queue sizes, leading to higher processing time. The

observed linear memory usage is reasonable for the current scale but may become super-linear with larger workloads (e.g., 100+ transactions), due to recursive depth growth, expanding BFS queues, and cumulative metadata storage for conflict tracking and priority management. The decline in throughput is linked not only to increased lock contention but also to architectural constraints in the simulation. MATLAB's single-threaded environment lacks parallelism and asynchronous processing, meaning lock requests are handled sequentially and transaction overlap is restricted—conditions that do not fully reflect real-world DBMS capabilities. As a result, throughput suffers from compounding delays caused by synchronous lock acquisition and conflict resolution.

While Tree-HP2PL introduces some computational overhead, particularly through priority inheritance and hierarchical traversal, it avoids cascading aborts and image storage overheads associated with protocols like 2PLNTHP and S-PROMPT. These comparisons are analytical, based on performance trends and behavior reported in prior literature [67], [68] rather than from direct experimental benchmarking. Overall, Tree-HP2PL demonstrates a strong balance of correctness, scalability, and responsiveness under increasing load in nested transaction environments. However, future work should focus on deploying the protocol in multi-threaded or distributed settings to validate its operational robustness and further optimize its performance in realistic real-time systems.

In addition to these empirical findings, the complexity and design of the Tree-HP2PL lock acquisition algorithm also warrant discussion. The algorithm combines Depth-First Search (DFS) and parallel Breadth-First Search (BFS) to manage lock conflicts in nested transactions. The worst-case time complexity of both DFS and BFS components is O (V + E), where V is the number of transactions and E is the number of transaction dependencies. DFS enables efficient deep traversal of hierarchical parent-child relationships to resolve intra-transaction conflicts, while parallel BFS handles inter-transaction or sibling-level conflicts concurrently. This combination enhances scalability by addressing both vertical and horizontal conflict patterns within transaction trees. However, as transaction volume increases, particularly beyond 20 transactions, traversal overhead and dynamic lock and priority state maintenance introduce execution time and memory overhead. These trends are evident in the simulation results, which show increased execution and response times at higher transaction volumes. While this trade-off impacts scalability, the DFS-BFS approach provides a deterministic, modular method of resolving conflicts—particularly beneficial in complex, real-time nested systems. In comparison, hybrid optimistic-pessimistic locking schemes may reduce overhead in low-contention environments but often suffer from validation delays and less predictable behavior under load. Tree-HP2PL's use of priority inheritance, though effective in mitigating priority inversion, introduces additional computational costs due to recursive priority propagation and increased memory usage. Nonetheless, this strategy ensures fairness and deadlock avoidance, which are critical for correctness in distributed real-time systems. Overall, Tree-HP2PL's algorithm design strikes a practical balance between responsiveness, correctness, and scalability, though further optimizations may be required for extremely high-throughput scenarios.

The Tree-HP2PL protocol was evaluated in a MATLAB simulation environment, chosen for its strengths in rapid algorithm prototyping, matrix manipulation, and visualization. While MATLAB does not replicate the full concurrency behavior of modern DBMSs like PostgreSQL or Oracle, it serves as a practical platform for modelling complex transaction hierarchies, simulating resource contention, and analyzing conflict resolution strategies in a controlled setting. One aspect influencing performance is the distribution of transaction priorities. Tree-HP2PL performs consistently under uniform or balanced priority distributions, with minimal aborts and stable throughput. However, when priorities are heavily skewed or dynamically shifting—such as a concentration of high-priority transactions—lock contention intensifies, and priority inversion risks increase. The protocol's integrated priority inheritance and selective abort mechanisms mitigate these effects to maintain fairness, but performance sensitivity to workload characteristics remains, suggesting a need for adaptive tuning in real-time systems. Beyond simple priority-based conflict resolution, Tree-HP2PL handles deadlocks through a dual detection strategy using DFS and parallel BFS to identify dependency cycles. This approach analyses both waits-for-lock and waits-for-commit graphs to detect and resolve complex deadlocks typical in nested environments. Under high-contention scenarios, where multiple transactions simultaneously request overlapping resources, the system escalates to priority inheritance and carefully chosen aborts to break deadlock chains. The use of parallel BFS distributes the conflict resolution load across sibling nodes, improving responsiveness. Nonetheless, performance degradation can occur if contention remains persistently high, underscoring the importance of incorporating resource-aware scheduling or load-throttling mechanisms in practical implementations. Overall, Tree-HP2PL demonstrates robustness across variable workloads and complex conflict scenarios, though its effectiveness can be further enhanced through integration with adaptive and distributed runtime strategies.

10 Future directions and challenges

This section identifies critical future areas in database management systems, focusing on advanced extended transaction models, adaptive concurrency control, and overhead optimization. It also emphasizes the potential of hybrid approaches, enhancements in distributed real-time systems, and the integration of machine learning for intelligent optimization, addressing the evolving needs of modern applications.

In the rapidly advancing field of database management systems, future directions and challenges play a pivotal role in shaping the landscape of technology. As we delve into the next phase of innovation, several key areas emerge where further exploration and development are essential to meet the evolving demands of modern applications and systems.

10.1 Extended transaction models

As we look ahead, an exciting avenue for research involves the exploration of more intricate extended transaction models. Tailoring these models for specific application domains, such as CAD/CAM and software engineering, can lead to heightened efficiency and improved support for specialized workflows. For instance, the Tree-HP2PL protocol could be enhanced with semantic-aware locking strategies, allowing it to adapt to domain-specific object hierarchies and dependencies in structured design systems.

10.2 Adaptive concurrency control

The future beckons the development of adaptive concurrency control mechanisms. These mechanisms would dynamically adjust based on transaction characteristics and workload patterns, ushering in a new era of responsiveness and performance optimization within database systems. A promising direction is to extend Tree-HP2PL with a feedback-based adaptation mechanism, where the system monitors runtime metrics (e.g., lock wait times or subtree density) and switches between DFS and parallel BFS strategies, accordingly, optimizing performance in real-time.

10.3 Optimizing overhead

Addressing the overhead associated with concurrency control protocols, especially in optimistic approaches, remains a pressing challenge. Future research efforts could concentrate on optimizing storage and processing requirements, paving the way for more streamlined and resource-efficient database systems. One approach could involve lock compaction, in which sibling subtransactions with similar lock access patterns share metadata, reducing redundancy and memory footprint without sacrificing isolation guarantees.

10.4 Hybrid concurrency control

In the quest for versatile and efficient solutions, there is a need to investigate hybrid approaches that combine the strengths of both optimistic and pessimistic concurrency control. Such hybrid models could offer a balanced approach, enhancing flexibility and performance across diverse application scenarios. For example, a hybrid version of Tree-HP2PL could be devised where critical parent-level transactions follow the pessimistic path, while leaf-level subtransactions execute under optimistic assumptions, validated upon commit.

10.5 Distributed real-time database systems

The evolution of distributed real-time database systems is critical in response to the increasing prevalence of distributed computing and the growing demands of real-time applications. Ongoing research in this domain should continue to refine and expand the capabilities of these systems to meet the challenges of an interconnected, real-time world. Enhancing Tree-HP2PL with deadline-aware scheduling algorithms—such as Earliest Deadline First (EDF) could significantly improve its responsiveness under time-constrained, distributed workloads.

10.6 Machine learning integration

The integration of machine learning techniques into database systems holds immense promise for the future. Exploring ways to leverage machine learning for adaptive concurrency control and predicting transaction behaviors could usher in a new era of intelligent, self-optimizing database systems, further enhancing their adaptability to dynamic workloads. Specifically, Tree-HP2PL could be extended with a reinforcement learning (RL) module trained to adjust transaction priorities based on system history and conflict likelihood, thereby preventing bottlenecks before they occur.

In summary, the future of database management systems hinges on addressing emerging challenges through innovative solutions, such as advanced transaction models, adaptive concurrency control, and the integration of machine learning. These developments are essential for enhancing system efficiency, responsiveness, and adaptability to meet the dynamic requirements of modern applications and industries. Continued research in these areas—especially through the development of semantic locking, adaptive traversal strategies, hybrid control models, and AI-powered prioritization—will pave the way for more robust and intelligent database systems.

11 Conclusion

This section emphasizes the Tree-HP2PL protocol's strong performance in nested transactions while identifying the need for better resource management to

mitigate increased execution and response times with higher transaction volumes.

This paper provides a comprehensive evaluation of the Tree-HP2PL concurrency control protocol within nested transactions, showcasing its exceptional performance across key metrics. The experimental results underscore the protocol's efficiency in managing crucial aspects such as system throughput, success ratio, response time, execution time, and memory usage within nested transaction scenarios. The analysis demonstrates that the performance of the Tree-HP2PL protocol is directly influenced by the number of transactions, with noticeable impacts on execution time, memory usage, system throughput, and response time. As the number of transactions increases, several key performance metrics experience notable changes. The average execution time shows a gradual rise, indicating increased processing overhead and potential bottlenecks, particularly evident at 20 transactions. Memory usage exhibits a linear growth pattern, directly correlating with transaction volume, highlighting the system's escalating resource demands. System throughput demonstrates a decrease, suggesting potential processing constraints under higher transaction volumes. Despite these challenges, the success ratio remains consistently high, showcasing the protocol's reliability in completing transactions successfully. However, response time increases with transaction volume, emphasizing the need for optimized system responsiveness. These findings underscore the importance of efficient resource management and conflict resolution mechanisms within the Tree-HP2PL protocol, especially when handling varying transaction workloads. Looking ahead, addressing these performance nuances, and exploring alternative algorithms can further enhance the protocol's scalability, efficiency, and overall performance in real-world transaction management scenarios.

Overall, the analysis of the Tree-HP2PL concurrency control protocol within nested transactions highlights its strong performance and reliability, reflected in consistently high success ratios. However, the increasing execution and response times at higher transaction volumes indicate a pressing need for improved resource management and optimization strategies. To enhance the protocol's scalability and efficiency, future research should focus on addressing these performance challenges and exploring alternative algorithms that can better accommodate varying transaction workloads in practical applications

References

[1] Connolly, T., & Begg, C. (1998). *Database systems: A practical approach to design, implementation and management* (2nd ed.). Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

[2] Özsu, M. T., & Valduriez, P. (2020). *Principles of distributed database systems* (4th ed.). Springer, pp. 1–674. https://doi.org/10.1007/978-3-030-26253-2.

[3] Ramamritham, K. (1993). Real-time databases. *Distributed and Parallel Databases*, 1(2), 199–226. https://doi.org/10.1007/BF01264051.

[4] Abbott, R., & Garcia-Molina, H. (1988). Scheduling real-time transactions: A performance evaluation. *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB)*.

[5] Stankovic, J. A., & Zhao, W. (1988). On real-time transactions. *Proceedings of the ACM SIGMOD Conference*, March 1988. https://doi.org/10.1145/44203.44204.

[6] Choudhary, M., & Haritsa, J. M. (1992). Data access scheduling in firm real-time database systems. *International Journal of Real-Time Systems*, 4(3). https://doi.org/10.1007/BF00365312.

[7] Mukkamala, M., & Shanker, U. S. (2008). Distributed real-time database systems: Background and literature review. *International Journal of Distributed and Parallel Databases*, 23(2), 127–149. https://doi.org/10.1007/s10619-008-7024-5.

[8] Gray, J. (1980). A transaction model. *Proceedings of the International Colloquium on Automata, Languages, and Programming (ICALP)*. https://doi.org/10.1007/3-540-10003-2_78.

[9] Gray, J. (1978). Notes on database operating systems. In *Operating Systems—An Advanced Course* (Lecture Notes in Computer Science, Vol. 60, pp. 393–481). Berlin: Springer-Verlag. https://doi.org/10.1007/3-540-08755-9_9.

[10] Gray, J. (1981). The recovery manager of the System R database manager. *ACM Computing Surveys*, 13, 223–244. https://doi.org/10.1145/356842.356847.

[11] Rastogi, S., Bhargava, H. K., & Silberschatz, A. (1993). On correctness of non-serializable executions. *Proceedings of the 12th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. https://doi.org/10.1145/153850.153859.

[12] Abiteboul, S., & Vianu, V. (1988). Equivalence and optimization of relational transactions. *Journal of the ACM*, 35, 70–120. https://doi.org/10.1145/42267.42271.

[13] Gray, J. (1987). Why do computers stop and what can be done about it. *CIPS (Canadian Information Processing Society) Edmonton '87 Conference Tutorial Notes*, Edmonton, Canada.

[14] Papadimitriou, C. H. (1979). Serializability of concurrent database updates. *Journal of the ACM*,

26(4), 631–653. https://doi.org/10.1145/322154.322158.

[15] Papadimitriou, C. H., Bernstein, P. A., Rothnie, J. B., & Stearns, R. E. (1976). Concurrency controls for database systems. *Proceedings of the 17th Symposium on Foundations of Computer Science (FOCS)*.

[16] Kung, H. T., & Papadimitriou, C. H. (1979). An optimality theory of concurrency control for databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/582095.582114.

[17] Carey, M. J., & Livny, M. (1988). Distributed concurrency control performance: A study of algorithm, distribution, and replication. *Proceedings of the 14th International Conference on Very Large Data Bases (VLDB)*, Los Angeles, California.

[18] Chung, Y., & Gruenwald, L. (1994). Research issues for a real-time nested transaction model. *Proceedings of the 2nd IEEE Workshop on Real-Time Applications*, July 1994.

[19] Medjahed, M. O., & Elmagarmid, A. K. (2009). Generalization of ACID properties. In *Encyclopedia of Database Systems*. Boston, MA: Springer. https://doi.org/10.1007/978-0-387-39940-9_736.

[20] Garcia-Molina, H., & Salem, K. (1987). Sagas. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/38714.38742.

[21] Rusinkiewicz, M., & Sheth, A. (1995). Specification and execution of transactional workflows. In W. Kim (Ed.), *Proceedings of the International Conference on Management of Data* (pp. 592–620). New York, NY: ACM Press/Addison-Wesley.

[22] Pu, C. (1988). Superdatabases for composition of heterogeneous databases. *Proceedings of the 4th International Conference on Data Engineering*. https://doi.org/10.1109/ICDE.1988.105502.

[23] Zaslavsky, A., & Bhargava, B. (2009). *Flex transactions*. In Ö. M. Liu, L. (Ed.), Boston, MA: Springer.

[24] Moss, J. E. B. (1981). *Nested transactions: An approach to reliable distributed computing*. Cambridge, MA.

[25] Weikum, G., & Schek, H. (1991). Multi-level transactions and open nested transactions. *Proceedings of the International Workshop on Database Systems for Advanced Applications*.

[26] Bjork, L. A. (1973). Recovery scenario for a DB/DC system. *Proceedings of the ACM Annual Conference*. https://doi.org/10.1145/800192.805695.

[27] Davies, C. T. (1973). Recovery semantics for a DB/DC system. *Proceedings of the ACM Annual Conference*. https://doi.org/10.1145/800192.805694.

[28] Guerraoui, R. (1995). Nested transactions: Reviewing the coherence contract. *Elsevier Science Journal*, 84, 161–172. https://doi.org/10.1016/0020-0255(94)00118-U.

[29] Karabatis, G. (2017). Nested transaction models. In *Encyclopedia of Database Systems*. New York, NY: Springer. https://doi.org/10.1007/978-1-4899-7993-3_716-2.

[30] Buchmann, A. (2016). Open nested transaction models. In *Encyclopedia of Database Systems*. New York, NY: Springer. https://doi.org/10.1007/978-1-4899-7993-3_717-2.

[31] El-Sayed, H. S. H., & Ahmed, M. E. E.-S. A. A. (2001). Effect of shaping characteristics on the performance of nested transactions. *Information and Software Technology*, 43(10), 579–590. https://doi.org/10.1016/S0950-5849(01)00164-1.

[32] Haerder, T., & Rothermel, K. (1993). Concurrency control issues in nested transactions. *VLDB Journal*, 2(1), 39–74. https://doi.org/10.1007/BF01231798.

[33] Bernstein, P. A., & Goodman, N. (1981). Concurrency control in distributed database systems. *ACM Computing Surveys*, 13(2), 185–222. https://doi.org/10.1145/356842.356846.

[34] Bernstein, P. A., Hadzilacos, V., & Goodman, N. (1987). *Concurrency control and recovery in database systems*. Reading, MA: Addison-Wesley.

[35] Thomas, R. H. (1979). A majority consensus approach to concurrency control for multiple copy databases. *ACM Transactions on Database Systems*. https://doi.org/10.1145/320071.320076.

[36] Kung, H. T., & Robinson, J. T. (1981). On optimistic methods for concurrency control. *ACM Transactions on Database Systems*, 6(2), 213–226. https://doi.org/10.1145/319566.319567.

[37] Eswaran, K. P., Gray, J. N., Lorie, R. A., & Traiger, I. L. (1976). The notions of consistency and predicate locks in a database system. *Communications of the ACM*, 19(11), 624–633. https://doi.org/10.1145/360363.360369.

[38] Alsberg, P. A., & Day, J. D. (1976). A principle for resilient sharing of distributed resources. *Proceedings of the 2nd International Conference on Software Engineering*.

[39] Stonebraker, M. (1979). Concurrency control and consistency of multiple copies of data in distributed INGRES. *IEEE Transactions on Software*

Engineering, 5(3), 188–194. https://doi.org/10.1109/TSE.1979.234180.

[40] Lindsay, B., Mohan, C., & Obermarck, R. (1986). Transaction management in the R* distributed database management system. *ACM Transactions on Database Systems*, 11(4), 378–396. https://doi.org/10.1145/7239.7266.

[41] Tandem Computers. (1987). NonStop SQL: A distributed high-performance, high-availability implementation of SQL. *Proceedings of the International Workshop on High-Performance Transaction Systems*.

[42] Tandem Computers. (1988). A benchmark of NonStop SQL on the debit credit transaction. *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

[43] Borr, A. (1988). High performance SQL through low-level system integration. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/971701.50244.

[44] Hehner, D. H., & Verjus, J. P. (1979). An algorithm for maintaining the consistency of multiple copies. *Proceedings of the 1st International Conference on Distributed Computing Systems*.

[45] Reed, D. P. (1978). *Naming and synchronization in a decentralized computer system*. Cambridge, MA.

[46] Reed, D. P. (1983). Implementing atomic actions on decentralized data. *ACM Transactions on Computer Systems*, 1, 3–23. https://doi.org/10.1145/357353.357355.

[47] Batory, D., & Mahbod, B. (1988). Generalized version control in an object-oriented database. *Proceedings of the IEEE 4th International Conference on Data Engineering*, February 1988.

[48] Chou, H. T., & Kim, W. (1986). A unifying framework for versions in a CAD environment. *Proceedings of the International Conference on Very Large Data Bases*, Kyoto, Japan.

[49] Chou, H. T., & Kim, W. (1988). Versions and change notification in an object-oriented database system. *Proceedings of the Design Automation Conference*. https://doi.org/10.1109/DAC.1988.14770.

[50] Kao, B., Lee, K. Y., & Chiu, A. L. (1997). Comparing two-phase locking and optimistic concurrency control protocols in multiprocessor real-time databases. *Proceedings of the 5th International Workshop on Parallel and Distributed Real-Time Systems and the 3rd Workshop on Object-Oriented Real-Time Systems*.

[51] Carey, M. J., & Livny, M. (1987). Concurrency control performance modeling: Alternatives and implications. *ACM Transactions on Database Systems*, December 1987.

[52] Ramamritham, K., & Stankovic, J. A. (1985). Evaluation of a flexible task scheduling algorithm for distributed hard real-time systems. *IEEE Transactions on Computers*, 34(12), 1130–1143. https://doi.org/10.1109/TC.1985.6312211.

[53] Peterson, J. L., & Silberschatz, A. (1985). *Operating system concepts*. Reading, MA: Addison-Wesley Publishing Company.

[54] Stankovic, J. A., & Huang, D. T. J. (1991). On using priority inheritance in real-time databases. *Proceedings of the Twelfth IEEE Real-Time Systems Symposium*.

[55] Davari, S., & Sha, L. (1992). Sources of unbounded priority inversion in real-time systems and a comparative study of possible solutions. *ACM Operating Systems Review*, 26(2), 110–120. https://doi.org/10.1145/142111.142126.

[56] Haque, W. U. (1993). *Transaction processing in real-time database systems*.

[57] Lynch, N. (1986). Concurrency control for resilient nested transactions. *Advances in Computing Research*, 3, 335–376.

[58] Lynch, N., & Merritt, M. (1986). *Introduction to the theory of nested transactions*. Theoretical Computer Science, 62, 123–185. https://doi.org/10.1016/0304-3975(86)90014-9.

[59] Fekete, A., Lynch, N., Merritt, M., & Weihl, W. E. (1987). Nested transactions and read/write locking. *Proceedings of the 6th ACM Symposium on Principles of Database Systems*, San Diego, CA. https://doi.org/10.1145/28659.28669.

[60] Fekete, A., & Lynch, N. (1991). Multi-granularity locking for nested transaction systems. *Proceedings of the MFDBS'91 Conference*.

[61] Fekete, A., Lynch, N., Merritt, M., & Weihl, W. E. (1990). Commutativity-based locking for nested transactions. *Journal of Systems Sciences*, 41, 65–156. https://doi.org/10.1016/0022-0000(90)90034-I.

[62] Madria, S. N., & Kumar, B. C. S. (1997). Formalization and correctness of a concurrency control algorithm for an open and safe nested transaction model using I/O automaton model. *Proceedings of the 8th International Conference on Management of Data (COMAD'97)*, Madras, India.

[63] Rundensteiner, E. A., & Harder, T. (1995). Concurrency control in nested transactions with enhanced lock modes for KBMSs. *Proceedings of the 6th International Conference on Database and Expert Systems Applications (DEXA'95)*, London, UK.

[64] Garcia-Molina, H., & Lynch, N. (1994). Nested transactions and quorum consensus. *ACM Transactions on Database Systems*, 19, 537–585. https://doi.org/10.1145/195664.195666.

[65] Fekete, A., & Kameda, T. (1989). Concurrency control of nested transactions accessing B-trees. *Proceedings of the 8th ACM Symposium on Principles of Database Systems*.

[66] Madria, S. N., & Kumar, B. C. S. (1997). Formalization of linear hash structures using nested transactions and I/O automaton model. *Proceedings of the IADT '98 Conference*, Berlin, Germany.

[67] Abdouli, A. M., & Abdouli, M. A. (2005). A system supporting nested transactions in DRTDBSs. *Proceedings of the 1st International High-Performance Computing Conference*, September 2005. https://doi.org/10.1007/11557654_99.

[68] Abdouli, M. A. (2005). Scheduling distributed real-time nested transactions. *Proceedings of the IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC)*. IEEE Computer Society. https://doi.org/10.1109/ISORC.2005.49.

[69] Huang, J., & Ramamritham, K. (2000). The prompt real-time commit protocol. *IEEE Transactions on Parallel and Distributed Systems*, 11(2), 160–181. https://doi.org/10.1109/71.841752.

[70] Resende, T. H. A. G., Furtado, J. L. R., & Resende, F. (1997). Detection arcs for deadlock management in nested transactions and their performance. In *Advances in Databases, BNCOD* (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer.

[71] Haque, W. U. (1993). *Transaction processing in real-time database systems*.

[72] Haerder, T., & Reuter, K. R. (1987). Concepts for transaction recovery in nested transactions. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/38714.38741.

[73] Awoyelu, I., & Osinuga, P. (2012). Mathlab implementation of quantum computation in searching an unstructured database. *Informatica (Slovenia)*, 36, 249–254.